Identification and Characterization of miRNA regulatory networks

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer. nat.)

im Fach Biologie

eingereicht an der

Lebenswissenschaftlichen Fakultät

der Humboldt-Universität zu Berlin

von

Andrei Filipchyk

Präsidentin der Humboldt-Universität zu Berlin

Prof. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät

Prof. Dr. Bernhard Grimm

Gutachter: 1. Nikolaus Rajewsky

2. Andreas Herrmann

3. Markus Landthaler

Tag der mündlichen Prüfung: 24.04.2018

Zusammenfassung

Post-transkriptionelle Genregulation ist ein zentraler Mechanismus, den lebende Organismen nutzen, um Funktionalität, Entwicklung und Anpassung zu gewährleisten. Defizite in diesem Mechanismus haben zahlreiche Krankheiten und Fehlfunktionen zur Folge. Post-transkriptionelle Genregulation wird von RNA-bindenden Proteinen (RBPs) ausgeführt. Ihr kombinatorisches Agieren ermöglicht eine genau abgestimmte Kontrolle räumlicher und zeitlicher Genexpression. Ein RBP erkennt seine Zielmoleküle typischerweise anhand sogenannter Bindemotive: Nukleotidsequenzen, die kompatibel sind mit einer Aminosäuretasche innerhalb des Proteins. Es gibt jedoch einen Sonderfall der Zielmolekülerkennung, der überRNAs, insbesondere microRNAs (miRNAs), vermittelt wird. miRNAs sind im Genom kodierte 20-25 Nukleotid lange RNAs, die in Argonaut (Ago)-Proteine geladen werden können, um diese zu ihren Zielmolekülen (z.B mRNAs) zu navigieren. Es wird angenommen, dass miRNA:Ago-Komplexe nahezu alle zellulären Prozesse kontrollieren. Dementsprechend werden miRNA-Fehlfunktionen (z.B. verursacht durch Mutation nur eines einzelnen Nukleotids in einer Bindestelle) mit zahlreichen Erkrankungen in Verbindung gebracht. Die Charakterisierung aller miRNA-Zielmoleküle („miRNA targetome“) ist eine der wichtigsten Fragen, die mithilfe der Systembiologie adressiert werden kann.

“Crosslinking and immunoprecipitation” (CLIP)-Methoden wurden angewandt um Ago-Bindestellen in einer Reihe relevanter biologischer Systeme experimentell zu identifizieren. Allerdings weisen diese Verfahren einer bestimmten Bindestelle nicht direkt die Identität der bindenden miRNA zu. Rein theoretischen Vorhersagen anhand von miRNA-Sequenzkomplementarität fehlt es andererseits sowohl an Spezifität als auch an Sensitivität, da die miRNA-Zielerkennung auf nur kurzen Sequenzmotiven (6-8 Nukleotide) beruht. Darüberhinaus basieren die theoretischen Vorhersagen auf bereits bekannten miRNA-Binderegeln und können dementsprechend unser Verständnis der Zielmolekülerkennung nicht erweitern. Obwohl die Kombination von CLIP mit rechnerbasierten Methoden die Vorhersage von miRNA:Zielgen-Interaktionen verbessert, bleibt die Spezifität ein Problem. Um die genannten Limitationen zu überwinden, modifizierten wir das CLIP-Protokoll durch Hinzufügen eines Ligationsschrittes zur Generierung von chimären miRNA:Zielmolekül-Sequenzen, die direkte miRNA:Zielmolekül-Interaktionen repräsentieren. Zusätzlich waren wir in der Lage miRNA:Zielgen-Interaktionen in unmodifizierten CLIP-Proben zu identifizieren, was durch eine interne Ligaseaktivität erklärt werden kann, die Eukaryoten möglicherweise gemein ist. Aus diesem Grund analysierten wir publizierte CLIP-Datensätze neu. Insgesamt identifizierten wir 40000 spezifische miRNA:Zielgen-Interaktionen für C. elegans, Maus, Mensch und virusinfizierte Zellen. Wir validierten unsere Zielgene mit einer Reihe von rechnerbasierten und experimentellen Tests. Darüberhinaus konnten wir durch Kenntnis der direkten Interaktionen erstmalig nicht-kanonische miRNA-Bindungsmodi umfangreich charakterisieren. Schließlich entschlüsselten wir ein faszinierendes regulatorisches Zusammenspiel im menschlichen Gehirn, welches zwei miRNAs, eine zirkuläre RNA und eine lange nicht-kodierende RNA involviert.

Abstract

Post-transcriptional gene regulation is a key mechanism exploited by living organisms to ensure their functionality, development and adaptation. Deficiencies in this mechanism lead to various diseases and malfunctions. Post-transcriptional gene regulation is exerted by RNA-binding proteins (RBPs). Their combinatorial action allows fine-tuned control over spatial and temporal gene expression to meet the actual cell demands. An RBP typically recognizes its targets via so called binding motifs: nucleotide sequences compatible with an amino-acid pocket inside the protein. However, there is a special case of target recognition guided by RNAs. In particular, micro RNAs(miRNAs) – 20-25 nucleotide long transcripts encoded in the genome–can be loaded into Argonaute (Ago) proteins to navigate them to their target RNAs. It is estimated that miRNA:Ago complexes control virtually all processes occurring in the cell. Consequently, malfunctions in the miRNA pathway (including even a single nucleotide mutation in a binding site) are implicated in multiple disorders. Therefore, the characterization of the "miRNA targetome" is one of the most important questions addressed to the systems biology.

Crosslinking and immunoprecipitation (CLIP) methods have been applied to experimentally discover Ago binding sites for a number of relevant biological systems. However, these approaches do not directly assign miRNA identity to a particular binding site. Computational predictions based solely on miRNA sequence complementarity, on the other hand, lack both specificity and sensitivity, since miRNAs recognize their targets via short motifs (6-8 nucleotides). Moreover, as predictions are based on already known miRNA binding rules, they cannot improve our understanding of miRNA targeting. While combination of CLIP and computational approaches improves miRNA:target predictions, specificity remains an issue. To overcome the aforementioned limitations, we modified the CLIP protocol by including a ligation step to generate miRNA:target chimeric sequences representing direct miRNA:target interactions. Furthermore, we were able to recover miRNA:target interactions from unmodified CLIP samples, which can be explained by an internal ligation activity potentially common to all eukaryotes. Therefore, we reanalyzed published CLIP datasets. In total, we identified 40 000 unique miRNA:target interactions for C. elegans, mouse, human and virus-infected cells. We proved the validity of our targets with a number of computational and experimental tests. Moreover, the information on the direct interactions allowed us to characterize for the first time non-canonical miRNA binding modes on a large scale. Finally, we deciphered an intriguing regulatory interplay in human brain involving two miRNAs, a circular RNA and a long non-coding RNA.

Contents

Acknowledgements 7
Abbreviations 8
Preface 10
1. Introduction
1.1 The primacy of information 11
1.2 Design of biology 11
1.3 Modular nature of RNA 15
1.4 Post-transcriptional administrators 18
1.5 Physiological relevance of post-transcriptional regulation 22
1.6 Methods of miRNA targets identification 23
1.7 Direct identification of miRNA:target interactions 25
2. Materials and Methods
2.1 Ligation iPAR-CLIP 27
2.2 Computational Pipeline 29
2.3 Downstream analysis of ligation products 32
2.4 Mutagenesis and reporter assays to test miRNA interactions 35
2.5 ChiFlex methods 40
3. Results
3.1 ALG-1 iPAR-CLIP reveals thousands of miRNA binding sites in C. elegans 45
3.2 Ligation and control samples contain miRNA:target chimeric reads 47
3.3 Discovered chimeras represent endogenous miRNA:target interactions 49
3.4 Thousands of interactions were hidden in published AGO-CLIP datasets 53
3.5 miRNA seed matches are selected in course of evolution 56
3.6 miRNA:target chimeras allow to distinguish miRNA family members 58
3.7 Discovered miRNA:target interactions arise from relevant regulatory events 59
3.8 Analysis of ligation products unambiguously revealed targets of viral miRNAs 66
4. Results (ChiFlex)
4.1 ChiFlex is a tool to discover miRNA:target interactions 70
4.2 ChiFlex performs interactions with a controlled specificity 75
4.3 Exploration of miRNA targeting in human brain 79
4.4 The analysis of chimeras revealed regulatory modules in mammalian brain 83
5. Discussion
5.1 PAR-CLIP experiments augmented with a ligation step are able to retrieve true endogenous miRNA:target interactions with a high specificity 89
5.2 miRNA:target chimeras can be generated in conventional AGO-CLIP experiment 90
5.3 Chimeric reads allow deeper understanding of miRNA binding rules 92
5.4 miRNAs in mammalian brain 94
5.5 Potential applications and limitations of chimera-based methods 96
6. Bibliography 100
7. List of publications 110
Selbstständigkeitserklärung 111

Acknowledgements

I would never be able to pass through the challenges of my Ph.D. period without an incredible people who helped me a lot.

My deep acknowledgements go to:

Nikolaus Rajewsky, for being an excellent scientist and mentor. Working together with Nikolaus broadened my intellectual scope and helped to see scientific problems from the perspectives unimaginable before. I also want to acknowledge Nikolaus for his support and patience, which I needed a lot.

Stefanie Grosswendt, the best teammate I ever had. Her patience to science and incredible diligence were a driving force of our project. She, indeed, introduced me to the system biology and explained multiple aspects of the experimental work.

Marcel Schilling, for a great number of fruitful discussions, vital contributions to the data analyses and amazingly precise shooting.

Filippos Klironomos, for the help with the revision process and collaboration in miRNA analyses.

Panagiotis Papavasileiou and Marta Rodriguez, for being not only nice and supportive colleagues, but also excellent advisers and consults.

Sebastian Mackowiak, Marvin Jens, Andranik Ivanov and Ana Elefsinioti, the people who supported me greatly in the begging and taught me bioinformatics.

Kathrin Theil, very nice and helpful person, who guided me through the variety of complications.

Benedikt Obermayer, for the great improvement of my writing skills

Eva Gottwein and Mark Manzano, for the validation experiments of the discovered miRNA interactions in virus-infected cells.

Margareta Herzog, for her contribution to the experimental part of the project

All my lab members, for the incredible working climate and valuable lessons

I want to acknowledge my beloved family for great support throughout my life. They are the key stone of my life, and be it forever.

Abbreviations

4SU	4-thiouridine
AGO	Argonaute
AU-rich	Adenylate-uridylate-rich
C. elegans	*Caenorhabditis elegans*
cDNA	complementary DNA
CDR1	cerebellar degeneration-related protein 1
CDR1-AS	CDR1 antisense RNA
CDS	coding sequence
CLASH	crosslinking, ligation, and sequencing of hybrids
CLIP	crosslinking and immunoprecipitation
CRISPR	clustered regularly interspaced short palindromic repeat system
DNA	deoxyribonucleic acid
dsRNA	double stranded ribonucleic acid
EBV	Epstein-Barr-Virus
E. coli	*Escherichia coli*
FDR	false discovery rate
GEO	Gene Expression Omnibus
GFP	Green fluorescent protein
GO	Gene Ontology
GSE	GEO Series
HITS-CLIP	high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation
HEK 293	human embryonic kidney 293
HuR	human antigen R
IgM	Immunoglobulin M
iCLIP	individual-nucleotide-resolution CLIP
iPAR-CLIP	*in vivo* PAR-CLIP
IP	immunoprecipitation
KSHV	Kaposi's sarcoma-associated herpes virus
LRG	Logic Rules Generator
miRBase	database archiving miRNA sequences and annotations
miRISC	miRNA-Induced Silencing Complex
miRNA	microRNA
miRNA ID	miRNA identifier for a gene from the miRbase database
miRTarBase	database of MicroRNA-Target Interactions
mRNA	messenger RNA
NGM	Nematode growth medium
PAR-CLIP	photo-activatable-ribonucleoside-enhanced crosslinking and immunoprecipitation
RBD	RNA-binding domain
RBP	RNA binding protein
RIP	RNA immunoprecipitation
RNA	ribonucleic acid
RNAi	RNA interference
RNP	ribonucleoprotein complex
rRNA	ribosomal RNA
SILAC	stable isotope labeling by amino acids in cell culture
SNP	single nucleotide polymorphism
TRPB	Tryptophan synthase beta chain protein
tRNA	transfer RNA
UTR	untranslated region

Preface

The first part of this project has been published as:

Stefanie Grosswendt, Andrei Filipchyk, Mark Manzano, Filippos Klironomos, Marcel Schilling, Margareta Herzog, Eva Gottwein, Nikolaus Rajewsky, "Unambiguous Identification of miRNA:Target Site Interactions by Different Types of Ligation Reactions" Molecular Cell, Volume 54, Issue 6, 19 June 2014, Pages 1042–1054

Individual contributions for the project are outlined below:

Stefanie Grosswendt designed and performed the ALG-1 iPAR-CLIP ligation and control experiments. Margareta Herzog contributed technical assistance. I designed the computational pipeline and performed the discovery and analysis of miRNA:target interactions. Stefanie Grosswendt, Nikolaus Rajewsky and I interpreted the data on miRNA chimeras. Marcel Schilling performed conservation analysis and contributed to the computational pipeline development. Filippos Klironomos compared chimeras with computational predictions. Eva Gottwein and Mark Manzano performed the validation experiments for viral miRNA interactions. Nikolaus Rajewsky supervised the workflow of the projects disclosed here.

1. Introduction

1.1 The primacy of information

Each living being on the Earth is a derivative of previous generations. Consequently all the aspects of life, from complex body plans to molecular machinery, have roots and reasons in historical prospective. The maintenance, propagation and transformation of life rely on the transmission of the information stored in specific molecules. Thus, fundamental biological problems can be approached from the perspective of continuous flow and evolution of information over the time. According to this paradigm each living organism is merely a message with an ultimate goal to survive. Even though this statement may be considered outraging, living beings indeed intend to save their information rather than their physical bodies. Altruism – the mechanism to sacrifice one's life to those sharing similar genetic content - is a perfect illustration of the information primacy.

The biological information is stored in high-order arrangements of DNA or RNA molecules. According to the second law of thermodynamics each ordered structure needs a continuous supply of energy (or negative entropy according to Schrödinger) to compensate increasing entropy. Consequently, each message has to use an elaborate apparatus to get energy and convert it into negative entropy. As this apparatus can correspond only to the information encoded in the message, there should be a mechanism of translation from the message to its supporting structures. The supporting structures also undergo thermodynamical pressure and may degrade in a time course. Therefore, a message can benefit from copying itself and then reconstructing all the required machinery anew. However, a precise copying of intact information cannot help to respond to a changing environment, while the adjustments in the information content may be beneficial. Thus, each molecular structure, body plan or behavioral pattern of any organism is a product of a long selection under constantly changing environment. These beneficial adaptations are encoded in a genome and may be shared across different species. Therefore, a nucleotide sequence conserved in multiple species may be related to a functionally relevant trait. The meaning of information primacy in biology was aptly summarized in the title of the essay by Theodosius Dobzhansky “Nothing in Biology Makes Sense Except in the Light of Evolution”.

1.2 Design of biology

As we discussed in the previous section, in order to survive genetic information needs elaborate supporting machinery. This machinery has to be designed in a way to encounter multiple challenges: acquiring and processing of energy and chemicals, keeping the integrity of a genetic material, copying itself and many others. If we reverse the logic, then each peculiarity of the design of a living organism

(it includes not only the body plans and cell architectures, but also behavioral patterns) can be explained via a rational reason. On the other hand, a living being can be considered as merely an ensemble of various molecules. These molecules don't understand reasons; they move and interact to each other according to the universal laws of physics. Therefore, biology can be also approached via fundamental laws of physics applied to the structures constituting living organisms. The idea to consider biological objects as very complex thermodynamic systems roots from the seminal treatise 'What Is Life?' written by Erwin Schrödinger.

Summarizing the aforementioned, a soup of molecules has to be self-organized in a high-order structure to assure survival and propagation of the encapsulated information. As the fully open systems eagerly reach thermodynamic equilibrium with an environment (that is, they lose their order), living organisms benefit from being enclosed. Therefore a cell, which interior is enclosed with a lipid membrane, can serve as a basic unit for a living organism and be living organism itself. However, stochastically moving molecules inside a cell also have to be organized in functionally relevant structures. This high-level organization can be achieved via the specificity of the intermolecular interactions. That is, molecules have to participate in only those interactions and chemical reactions which are required by a cell. Since chemical reactions and interaction are generally very specific for the molecules with extensive secondary and tertiary structures, a cell can benefit from using them. Thus, a highly-ordered molecular system has to be encapsulated and rely on very specific chemical interactions assured by molecules with an elaborate structure.

As a living organism has to perform multiple tasks, it is rational to use various types of molecules, each being the most suitable for a particular purpose. However, life likely appeared as a random event, rather was engineered. The chance, that one type of molecules acquires suboptimal functionality is infinitely higher, than that multiple types of already specialized molecules meet each other to initiate the advent of life. It seems that this first type of molecule was RNA (the concept of "RNA world" was introduced by Francis Crick, Leslie Orgel and Carl Woese in 1960s). Indeed, RNA molecules can store information via sequence of nucleotides. Exact base-pairing between RNA molecules may allow them to be copied precisely. Moreover, even in the modern organisms some RNAs possess catalytic activities. Thus RNAs have a potential to store information, copy it and generate a flow of negative entropy and required chemicals. The only competence RNA cannot provide is structural protection and segregation from outer world. Luckily, fatty acids eagerly form lipid bilayers in aqueous environment forming distinct compartments.

However, the design based solely on RNAs has some disadvantages. (1) An RNA molecule should

stay unchanged to keep its information content intact. Consequently, it should avoid interactions with other molecules. On the other hand RNA in this design is an enzyme. It has to be involved in multiple chemical reactions, which interfere to its function as information carrier. Therefore it is favorable for an organism to divide information storage and enzymatic activity tasks between different types of molecules. It turns out that long polymers of amino acids (called proteins) are able to perform a wide range of metabolic exercises with a high specificity. Moreover, high specificity of protein:protein interactions allows these molecules to form high-order structures together with lipids enhancing mechanistic support. (2) Similar to RNA, DNA molecules can store information in the nucleotide sequences. As, compared to RNA, DNA molecules eagerly adopt stable double-helix structures under normal conditions, they are more suitable to secure long-term storage of the genetic information. Thus, DNA molecules are proper information carriers, while proteins are convenient for metabolic and structural functions. Consequently, RNA occurs obsolete, since it is inferior to DNA as a genetic material and has narrower range of enzymatic activities than proteins. Here we come to a confusing question: “Why does live need RNA?” Indeed, many years of evolution shaped current biological systems to be very efficient in the cognate environments. Hence RNA should be discarded and completely replaced by DNA like gas lighting was replaced by incandescent lamp. However, for living organisms encountered so far holds the schema, known as central dogma of molecular biology: genetic information is stored in DNA, particular foci on DNA, known as genes are transcribed into RNA molecules, and RNA molecules are then translated into proteins (Fig. i1). Would it be more robust and efficient to derive proteins directly from DNA? Probably yes, but the underside of this robustness is rigidity. Rigid systems fail to adapt and specialize, and hence don’t survive in changing and competitive environment. For the living organisms the adaptation, specialization and signaling are substantially assured by post-transcriptional gene regulation exerted on RNAs.

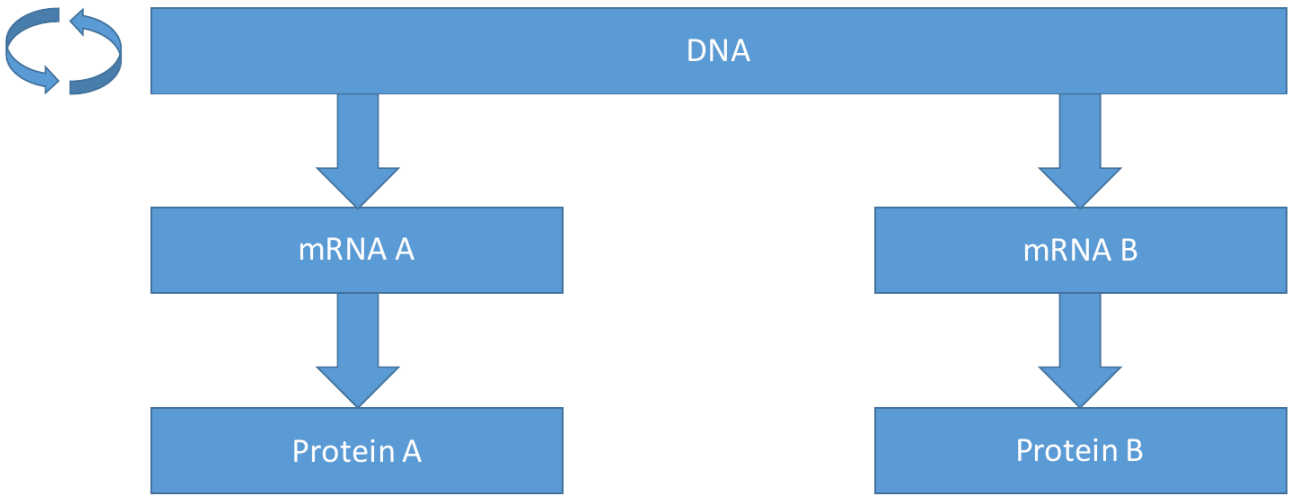

Figure i1| Central dogma of molecular biology
Nucleotide sequences are transcribed from DNA to RNA and the translated into proteins. DNA is duplicated upon cell division.

A cell in general requires constant genetic capacity and constant amount of membranes, with an exceptional case of cell division, when DNA and lipid molecules are doubled. Consequently, DNA and lipids are functionally static elements of a biological system. In opposite, the amounts of RNAs and proteins can be varied to establish a cell state suitable for particular functional tasks. For example, different protein expression patterns determine different specializations of liver and heart cells. Furthermore, a cell can adjust its protein content during differentiation, cell cycle, stress, and in response to the extracellular signaling. Thus, a precise regulation of protein levels is crucial for a cell functionality and communication with the environment.

There are three main routes to regulate protein expression: via control on RNA transcription from DNA, via control on RNA decay and translation, via control on protein decay. The regulation of transcription relies on a set of proteins known as transcription factors. They bind a stretch of DNA spatially close to the regulated genes enhancing or repressing their transcription. This fundamental idea of gene regulation was coined by Jacob and Monod in 1961 for the Lac operon in bacteria. It is advantageous to control gene expression on transcriptional level, since it saves energy and resources otherwise spent on unnecessary RNA production. However, transcriptional regulation is not flexible: it cannot adjust local concentrations of RNAs and proteins. Moreover for the cells with complex structure a switch in transcription may substantially lag the corresponding extracellular signal. For example: signal molecules have to travel about a meter along the axon of a neuron to reach the nuclei, enter it and change transcription rates. Then, newly synthesized RNAs and proteins have to travel all the way back. In opposite, spatial regulation can be performed on local levels of proteins or RNAs.

Post-transcriptional regulation (that is, exerted on RNAs) may be favorable comparing to the post-translational (that is, exerted on proteins) for the following reasons. (1) Since multiple protein molecules can be produced from one RNA transcript, the repression on RNA level leverages the repression on protein level. (2) RNAs are convenient objects for regulation, as they do not harbor any functional site. It means that regulatory agents can bind virtually all spots on RNA in combinatorial manner. In contrast, it is hard to imagine functional cis-regulatory site inside the tertiary structure of a protein. (3) It is also energetically beneficial to destroy RNA, rather than the cognate protein. Indeed, post-translational regulation mainly relies on a variety of protein modifications including phosphorylation, biotinylation, ubiquitination and others. (4) As genes are encoded on DNA in several distinct blocks (exons), multiple RNA transcripts may arise via alternative splicing of these exons. In contrast, for proteins it would be hard to imagine, that polypeptides produced on different ribosomes can be sewn in a controlled manner. Thus, post-transcription regulation is an energy efficient way to exert a fine-tuning combinatorial control on spatial and temporal gene expression with a significant leverage.

1.3 Modular nature of RNA

As an object of regulation, typical messenger RNA (mRNA) transcript consists of five functionally distinct regions (Fig. i2).

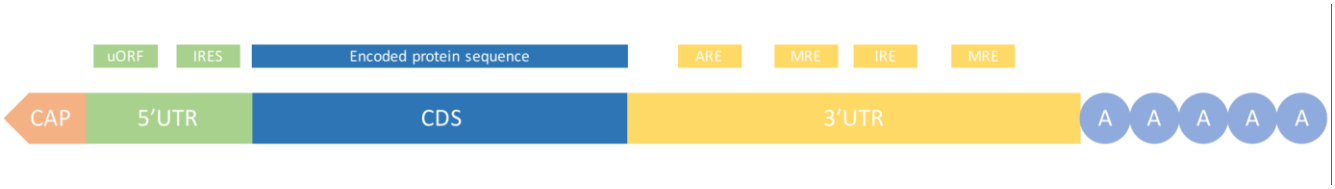

Figure i2| Modular composition of mRNA
Protein sequence is encoded by nucleotides in CDS. 3'UTR is a hub for regulatory sites. 5'UTR harbors elements involved in translation. PolyA tail and 5' cap assure transcript stability.

Coding sequence region (CDS): a part of mRNA translated into protein. As each 3 nucleotides (codon) in CDS are translated into one amino-acid, its sequence is under evolutionary selection for the amino acid sequence of the cognate protein. Consequently, there is an additional constrain on a potential cis-regulatory site to appear on CDS. This site has to be both sequence specific for the corresponding binding factor and comply with the codon structure. Moreover, CDS is not accessible for binding factors during translation. However, a number of cis-regulatory motifs in CDS were reported (Brümmer and Hausser, 2014; Liu et al., 2015). Potential regulatory sequences on CDS can be computationally predicted via analyses of the conservation of the third codon nucleotides, since the majority of amino acids are encoded by two first nucleotides, while the third may vary.

3' untranslated region (3'UTR): a part of mRNA downstream (that is, transcribed after) CDS. 3'UTR harbors multiple cis-regulatory sites including: adenylate-uridylate-rich elements (AREs), micro RNA (miRNA) response elements (MREs), iron response elements (IREs) and many others.

AREs are nucleotide stretches 50-150 nucleotides long strongly enriched with adenines and uridines. They are found in 5-8% of human genes (Halees et al., 2008). AREs are key determinants of transcript regulation. Depending on available binding factors they contribute to the decay or stabilization of their host transcripts. There are three main types of AREs' binding proteins: AUF-1, TTP and HU family. Combinatorial binding of these factors determines the fate of the targeted transcript. It was shown that AREs are involved in many pathological processes, including cancer (Hitti et al., 2016) and inflammation (Khabar, 2010).

MRE occur rather frequently throughout 3'UTRs. It is estimated that more than 60% of human genes carry miRNA binding sites (Friedman et al., 2009). Typically MRE is a sequence of 6-8 nucleotides, which is reverse-complement to a part of miRNA (nucleotides 2-7, 2-8 or 2-9) called seed. However many MREs require certain sequence context to be accessible for binding. Therefore the actual binding site may be longer. Since the number of miRNAs expressed in a typical cell type exceeds one hundred, multiple MREs allow precise combinatorial regulation of their host transcript.

IREs are cis-regulatory elements involved in iron metabolism. They are bound by factors dependent on cellular concentrations of iron. For example, a transcript of Ferritin, protein storing iron inside a cell, has IRE. In contrast to MREs and AREs, binding to IREs is determined by structure rather than sequence specificity. Indeed, IRE binding factors primarily recognize its stem-loop structure. Some of the nucleotide pairs forming the stem can be either G-C or A-U with no impact on binding specificity, while the sequence in the loop cannot be twisted. Thus, IRE is an example of

recognition based on both primary and secondary structure of mRNA.

Since 3'UTR is a hot spot for post-transcriptional gene regulation, its cis-regulatory sites repertoire may be dependent on host transcript function and cell context. Indeed many of transcriptional and post-transcriptional regulators have long 3'UTRs, while highly-expressed housekeeping genes use shorter ones (Stark et al., 2005). Even one gene may have multiple 3'UTR isoforms, opening a possibility for differential regulation in different cell types. For example, neuronal transcripts tend to have long 3'UTR isoforms compare to other tissues. 3'UTR isoform choice was shown to affect localization of the transcripts in neurons. Longer isoforms tend to reside in neuronal projections while, shorter isoforms accumulate in a cell body (Taliaferro et al., 2016). Moreover, 3'UTR may affect not only the localization of its transcript in a neuron, but also the localization of corresponding protein (Berkovits and Mayr, 2015). Thus, 3'UTR is a part of mRNA dedicated for combinatorial regulation via multiple binding factors, and is an object of regulation itself.

5' untranslated region (5'UTR): a part of mRNA upstream (that is, transcribed before) CDS. 5'UTR is known to harbor elements controlling translation. Upstream open reading frames (uORF) are present in around 50% of human protein coding genes. A uORF serves as a sponge for translational machinery, lowering protein production (Barbosa et al., 2013; Calvo et al., 2009). Furthermore, uORF may be also translated into a small peptide, which may be engaged in the interactions with the host transcript (Bazzini et al., 2014). 5'UTR may also contain internal ribosome entry site (IRES). IRES allows cap-independent translation, via recruiting ribosomes directly to the start codon. This mechanism is used by a number of viruses to avoid a check-up for the 5'cap by their host cells. Remarkably, 10-15 percent of mammalian protein coding genes were predicted to exploit IRES mechanism (Spriggs et al., 2008).

5'UTR may control translation rate independently on cis-regulatory sites. Relatively high GC content of a typical 5'UTR sequence in eukaryotes leads to the enlarged secondary structure (Leppek et al., 2017; Pesole et al., 1997), which prevents entry of a translational machinery (Babendure et al., 2006). Thus, a substantial fraction of 5'UTRs has a general property to inhibit translation. In contrast to 3'UTRs, 5'UTRs do not harbor plenty of binding-sites for trans-acting factors. So far only a few regulatory proteins were proved to have functionally relevant binding sites on 5'UTRs of their targets (Fred et al., 2016; Jungkamp et al., 2011).

Poly(A) tail: is a stretch of adenosines added to the 3' end of mRNA during transcription. Poly(A)

tail serves as an attractor for poly(A)-binding-protein. This interaction protects the transcript from degradation and facilitates mRNA export from nucleus to cytosol. mRNA stability and translation rate were reported to depend on poly(A) tail (Gorgoni and Gray, 2004; Meyer et al., 2004). Thus, gene expression can be regulated via alternative polyadenylation (APA). Indeed, 50-70% of human transcripts harbor multiple cleavage sites (Derti et al., 2012; Tian et al., 2005). Alternative polyadenylation relies on differential deadenylation controlled by trans-acting RBPs and miRNAs (Batra et al., 2015). Despite poly(A) tail plays an important role in translation, certain mRNA species, like histone protein coding transcripts, lack poly(A)-tail, but are still translated.

5' cap: a guanine nucleotide linked to the very 5' end of mRNA via triphosphate bridge. 5'cap serves as mark of quality for a transcript. It protects host mRNA from degradation and promotes translation. 5'cap may be removed during post-transcriptional regulation causing subsequent degradation of the uncapped mRNA.

In summary, mRNAs harbors multiple binding sites for various regulatory factors mainly residing on its 3'UTR. The temporal and spatial composition of the available factors determines essential aspects of mRNA life such as: degradation, localization, translation efficiency, and many others. Thus, local and temporal expression of mRNAs can be fine-tuned according to the actual cell demands.

1.4 Post-transcriptional administrators

Post-transcriptional gene regulation can be administrated by two major types of macromolecules: proteins and RNAs.

RNA-binding proteins (RBPs): proteins with high affinity to single- or double-stranded RNA molecules. High affinity is achieved via RNA-binding domains (RBD) – core components of RBPs. There are the following major types of RBDs: zinc fingers, K-homology domain, RNA recognition motif, heterogeneous nuclear RNP and double-stranded RNA-binding motif (Lunde et al., 2007; Mackereth and Sattler, 2012). However, according to the latest estimations only a half of around 900 of human RBPs have at least one known RNA-binding domain. RBPs recognize their targets via specific RNA primary or secondary structure. For example HuR, an RBP stabilizing targeted transcripts, binds to AU-rich sequences (López de Silanes et al., 2004; Soller and White, 2005). In opposite Microprocessor Complex, a protein complex involved in miRNA biogenesis recognizes hairpin structure of its targets, regardless to the sequence of a hairpin (Denli et al., 2004; Landthaler et al., 2004). Furthermore, there are RBPs with an affinity determined by both RNA sequence and structure. Vts1P, RBP found in yeast, binds CNGG motif within a loop of RNA hairpin (Aviv et al.,

2006). Similar to HuR, many RNA-binding proteins have loosely defined sequence preferences. A typical RBP binds a set of similar sequences with comparable affinities. This set is coined as a motif: sequence of nucleotides with the wildcards. For example any nucleotide can be at the second position of Vts1p CNGG motif. However, there is a very specific case of RNA-binding proteins which can recognize very different target sequences.

The members of Argonaute (Ago) family of endonculeases recruit a short (20-35 nucleotide) RNA. This small RNA (sRNA) guides Ago to the target sites via sequence complementarity, determining its specificity. As multiple sRNA species can be recruited, Ago has a very broad range of potential targets.
Together with guide RNA, Dicer and TRPB proteins Ago forms RNA-induced Silencing Complex (RISC), a key component of the phenomenon coined RNA-interference.

RNA interference (RNAi): a phenomenon of post-transcriptional gene regulation involving small non-coding RNAs. As it was mentioned above, gene regulation is a key mechanism of any self-organized biological system. Therefore, RNAi may root from the ancient 'RNA world' where RBPs were not available. Alternatively, RNAi can serve as a defensive weapon against invading genetic material (e.g. viruses), and hence it might acquire regulatory function as an addition to the protective one. RNAi is widespread among eukaryotes However some of unicellular eukaryotes and fungi, including model organism Saccharomyces cerevisiae, lack this mechanism (Aravind et al., 2000). Phylogenic analysis suggests that RNAi was lost in these organisms rather than never derived (Cerutti and Casas-Mollano, 2006). Prokaryotes do not have homologues to the components of eukaryotic RNAi. However clustered regularly interspaced short palindromic repeat (CRISPR) system, found in bacteria and archaea, provides an immunity against foreign genetic material similar to RNAi (van der Oost et al., 2009).

RNAi is initiated when endoribonuclease Dicer excises a short (20-30 nucleotides) fragment from a double-stranded RNA molecule. Each fragment is composed of passenger and guide strand. Guide strand is downloaded into an Argonaute protein, while passenger strand is degraded. The choice of a guide strand is independent on Dicer cleavage direction. It was shown that the strand of an RNA duplex with looser pairing at 5'end has more chances to be incorporated into Ago (Khvorova et al., 2003; Schwarz et al., 2003). An incorporated guide RNAs lays in a groove inside Ago, while it's 5'end is anchored to a binding pocket composed of conserved amino acids (Ma et al., 2005). Ago protein shapes the downloaded small RNA (sRNA) in a way to facilitate its basepairing with a targeted

transcript (Schirle et al., 2014).

Double-stranded RNA processed by Dicer may originate exogenously or endogenously. In the first case foreign genetic material is cut into pieces and loaded to Ago. These pieces are called small interfering RNA (siRNA). A siRNA guides Ago to the target RNAs via full sequence complementarity. That is, an Ago binding site is a reverse-complement to the guide siRNA. The length of a typical siRNA ranges from 20 to 25 nucleotide, which allows the detection of the invading genetic material with high specificity. The interaction between Ago-siRNA complex and an RNA transcript leads to the cleavage and subsequent degradation of the latter. Since single siRNA can be used many times to destroy its target, representation of a small amount of double-stranded RNA to a cell can cause significant consequences. Moreover, some species have mechanisms to leverage RNAi effects. For example, in model organism C. elegans RNAi can be amplified via synthesis of secondary siRNAs. They are produced by RNA-dependent RNA polymerase using primary siRNAs as templates (Pak and Fire, 2007). As RNAi can cause significant and specific gene silencing starting with a small amount of the represented double stranded RNA (dsRNA), it appears useful for a variety of biological applications, including functional genomics studies (Kamath et al., 2003).

A guide RNA for RISC complex can also have an endogenous origin. Micro RNA (miRNA) genes constitute a substantial fraction of non-coding transcriptome in animals, plants and some viruses. They typically reside in intergenic or intronic regions of a genome. However, there are few cases of miRNA transcripts originating from exons of coding genes. For example MIR-671 resides in the coding sequence of chondroitin polymerizing factor (CHPF2 in human). With a few exceptions miRNA genes are transcribed by polymerase II (Lee et al., 2004). As mRNAs, they are also co-transcriptionally capped and polyadenylated (Aukerman and Sakai, 2003; Cai et al., 2004). MiRNA primary transcripts (pri-miRNA) adopt a specific secondary structure with at least one long hairpin (around 70 nucleotides long). Hairpin structure is further recognized by nuclear RBP DGCR8. DGCR8 recruits a ribonuclease Drosha, which excises the hairpin from primary transcript. Since pri-miRNA may contain multiple hairpins, several different miRNAs can be produced from one polycistronic miRNA gene. For example human gene MIR17HG, known as mir17-92 cluster, encodes six different miRNAs. Co-expression of these miRNAs allows coordinated regulation of Bim protein, a gene involved in oncogenesis in B cells (Ventura et al., 2008). Cleavage by Drosha leaves two-nucleotide overhang at hairpin's 3' end, which is further recognized by Exportin-5 to ensure export from nucleus to cytoplasm (Yi et al., 2003). Some miRNA transcripts bypass processing by Drosha and are excised as hairpins directly from introns (Ruby et al., 2007). In the cytoplasm miRNA hairpins (pre-miRNA) are processed by endoribonuclease Dicer (Hutvágner et al., 2001; Park et al., 2011). It cuts a hairpin loop

out and leaves an imperfect duplex of 20-25 nucleotides. At this point miRNA biogenesis pathway converges with siRNA pathway. That is, one part of the duplex (guide miRNA) is loaded into Ago, while another (star miRNA) is degraded (Hammond et al., 2000).

Similar to siRNAs, plant miRNAs guide RISC to the transcripts with almost perfect complementarity, causing a cleavage of the targeted sequence. In animals miRNAs find their targets via partial complementarity, typically to the nucleotides 2-8 (Bartel, 2009). This recognition module on a miRNA is called 'seed', while its reverse-complement on a target sequence is called 'seed match'. However, a number of the discovered functional binding sites do not have a seed match for the cognate miRNA. Instead, these 'seedless' interactions can involve miRNA 3'end and/or imperfect (with several bulges) seed-match (Chi et al., 2012; Moore et al., 2015). Moreover, binding via the central region of human miR-124 was shown to repress Raptor gene (Shin et al., 2010).

In animals miRNA binding may lead to different functional outcomes. Targeted genes were shown to be cleaved, destabilized, repressed translationally or even activated. The mode of miRNA action depends on the architecture of binding sites and additional cofactors, but the exact mechanisms are not fully clear. However, the following schemes were proposed and investigated. (1) A transcript perfectly paired to a miRNA is cleaved by an Argonaute protein in front of miRNA nucleotides 10 and 11. (2) MiRNA induced silencing complex (miRISC) recruits CCR-NOT complex to the bound mRNA via GW182. CCR-NOT deadenilates mRNA, rendering it for the subsequent degradation. (3) MiRISC together with CCR-NOT may block the initiation of translational machinery on targeted transcript. (4) MiRISC can also affect translation inducing ribosome drop-off. (5) Argonaute in complex with fragile X mental retardation protein 1 (FMR1) may even promote an expression of the targeted transcripts (Mortensen et al., 2011). (6) MiRNAs were shown to bind DNA causing chromatin modifications followed by changes in the transcription rates (Catalanotto et al., 2016; Place et al., 2008). Thus, functional outcome of miRISC targeting depends on the downloaded miRNA, architecture of the binding site, available cofactors and cellular localization. However, miRNAs generally repress the expression of the targeted genes.

1.5 Physiological relevance of post-transcriptional regulation

In the previous sections we were discussing that post-transcriptional regulation is a convenient mean for a biological system to specialize, self-maintain and react to external signals. Indeed, it is vigorously used by the living organisms. For example, in human at least 800 genes code for RNA binding proteins (Baltz et al., 2012; Castello et al., 2012) and 2-5 thousands for miRNAs (Friedländer et al., 2012; Londin et al., 2015) constituting a substantial fraction of the whole transcriptome.

miRNAs were shown to regulate the expression of around 60% of protein coding genes (Friedman et al., 2009), being involved in control over almost all the processes which take place within a cell, including metabolism, development and signaling (Krol et al., 2010).

Unsurprisingly, deficiencies in post-transcriptional regulation affect the normal functionality of an organism, and are associated with a number of diseases (Hesse and Arenz, 2014; Sayed and Abdellatif, 2011). For instance, conditional knock-out of Dicer causes impaired differentiation of mouse embryonic stem cells (Kanellopoulou et al., 2005). Ablation of Argonaute, another key component of miRNA machinery, leads to the apoptosis of embryonic stem cells (Hong et al., 2009).

Apart from a global perturbation of miRNA pathway, a mis-regulation of a single miRNA may induce large-scale functional changes. For instance, impaired angiogenesis (which can be lethal) was initially associated with a loss of Dicer (Yang et al., 2005). However, further experiments revealed a specific role of miR-126 in the development of vascular system. Thus, the ablation of Dicer leads to the stunted biogenesis of miR-126 and consequently affects angiogenesis. There are numerous other examples of pathologically relevant miRNAs including: miR-1 controlling muscle development, which loss leads to the heart failure in mice (Sokol and Ambros, 2005); miR-155 involved in the immune response, which loss is associated with the oppressed production of IgM (Rodriguez et al., 2007). miR-278 playing an important role in energy homeostasis, as fruit flies lacking it have elevated insulin levels (Teleman et al., 2006). Furthermore, the interruption of a particular miRNA:mRNA interaction may lead to significant distortions. For example, single nucleotide polymorphisms (SNP) in binding sites for miR-204 and miR-211 on the caprine *KITLG* 3′-UTR affect the litter size (An et al., 2012). An SNP influencing the interaction between TYRP1 gene and miR-155 can serve as a melanogenic prognostic marker (El Hajj et al., 2015). The latter example also represents a considerable involvement of miRNAs in oncogenic processes. Indeed, a significant fraction of miRNA genes is located in cancer-associated genomic regions (Zhang et al., 2006), and some of them can be even lost during chromosomal rearrangements (Calin et al., 2002). MiRNAs can be both oncopromoters and oncosuppressors. MiR-15 and miR-16 were shown to induce apoptosis of the malignant cells (Cimmino et al., 2005). Let-7 miRNA also suppresses cancer development via translational inhibition of RAS and MYC oncogenes (Johnson et al., 2005; Takamizawa et al., 2004). In contrast, the overexpression of miRNAs from 17-92 cluster can cause leukemia (Jin et al., 2013; Mi et al., 2010; Sandhu et al., 2013). Thus, multiple evidences suggest an important role of miRNAs in development, metabolism and pathology. These roles directly depend on a set and composition of the available target genes. Consequently, the deciphering of miRNA:gene interaction map may contribute greatly to the understanding of the most relevant biological processes.

1.6 Methods of miRNA targets identification

As miRNAs typically downregulate targeted genes, the knock-down or a knock-out of a particular miRNA should cause derepression of its targets. Conversely, an overexpression of a miRNA leads to the inhibition of the targeted genes. Thus, if one perturbs expression levels of a miRNA and counts the changes in gene expression, then genes with significant changes can be considered as the targets for this miRNA. This scheme gives rise to a branch of methods coined here as "miRNA perturbation experiments". The perturbation approaches vary in how the miRNA expression is manipulated (genomic knock-outs, LNA, antagomirs, overexpression vectors, etc.) and how gene levels are detected (microarrays, RNA sequencing, GFP reporters). Despite being rather straightforward, perturbation methods have several disadvantages. (1) As miRNAs may target transcription factors and RNA-binding proteins, both direct and indirect miRNA targets may be discovered. (2) Perturbation experiments consume large amounts of time and resources; hence can be applied to characterize targets only for several miRNAs in a particular biological system. Thus, miRNA perturbation approach cannot be used to find direct targets for all the expressed miRNAs for a broad range of cell types. However, the databases harboring experimentally validated targets (Chou et al., 2016; Sethupathy et al., 2006) are found to be useful in miRNA research.

A number of computational tools to predict miRNA binding sites were developed in recent decades, including PicTar (Krek et al., 2005) and TargetScan (Lewis et al., 2005). They typically focus on search for miRNA seed-matches. As seed recognition motif is relatively short (6-8 nucleotides), it appears rather often in a genome. Consequently the predictions based solely on seed-matches have a high false discovery rate (FDR). In order to increase the specificity, the following attributes of potential miRNA binding sites are also taken into account: sequence conservation, structural accessibility, hybridization energy (with a cognate miRNA), and location on a gene (3'UTR binding sites are preferred). Even though combining these attributes into a single score improves the quality of predictions, the number of false discoveries remains an issue. Moreover, the pairing via seed site is not the only binding mode utilized by miRNAs. Consequently a significant fraction of non-canonical (that is, lacking perfect seed complementarity) binding sites is missed in the predictions, affecting also the sensitivity of computational methods. Indeed, an information content of a seed-match with one mismatch is low enough to render reliable predictions of such binding sites almost impossible. Thus, computational methods lack both specificity and sensitivity in the discovery of miRNA targets. Moreover, bioinformatics predictions do not address context-depended aspects of miRNA binding, such as competition for the binding sites with other RBPs and co-localization of a miRISC with its

targets.

Biochemical methods are also employed to decipher miRNA regulatory networks. They are based on an immunoprecipitation of a miRISC:target complex followed by the detection of a targeted RNA via microarrays or next-generation sequencing. The specificity and sensitivity of the pull-down of miRNA targets can be enhanced with UV irradiation, as it triggers a covalent linkage between an Argonaute protein and bound RNA. For instance, in HITS-CLIP (High-Throughput Sequencing Cross-Linking and Immuno Precipitation) cell cultures or even living animals are irradiated with a light of 365 nm wavelength (Chi et al., 2009). Furthermore, the crosslinked nucleotide may be incorrectly amplified during reverse-transcription, and hence mapped as a mismatch, deletion or insertion. Since the crosslink is favored to be formed right upstream the seed match, these substitutions can be used to achieve a sub-nucleotide resolution of miRNA binding sites detection. This idea was even further emphasized in iCLIP (individual-nucleotide resolution Cross-Linking and Immuno Precipitation) and PAR-CLIP (Photoactivatable Ribonucleoside–Enhanced Cross-Linking and Immuno Precipitation). In iCLIP reverse-transcription stops at the crosslinked nucleotide during amplification process, and hence sequencing reads are directly flanked with the crosslinks (Broughton and Pasquinelli, 2013; König et al., 2012). In PAR-CLIP photoreactive ribonucleoside analogs are incorporated into newly synthesized RNA (Hafner et al., 2010). Crosslink involving these artificial nucleotides causes T->C (or A->G, depending on a type of ribonucleoside) substitution in the following RNA amplification.

As CLIP methods do not directly assign miRNA identities to the binding sites, they are typically augmented with bioinformatics analysis. Predictions made on context-depended and narrow set of Argonaute binding sites are clearly superior to those made without experimental support. Moreover, several tools were developed to specifically benefit from substitutions introduced by crosslinks (Corcoran et al., 2011; Kerpedjiev et al., 2014). Even though combinations of CLIP methods and computational predictions perform with acceptable specificity, two major problems were encountered. (1) A considerable fraction of Argonaute binding sites lacks a seed match to any expressed miRNA. (2) Many miRNAs are arranged in miRNA families. That is, they have common seed sequence. Consequently computational predictions cannot distinguish between them. Thus, despite CLIP methods and bioinformatics tools applied together greatly improve the detection of miRNA targets, direct and reliable assignment of the miRNA identity remained unachievable.

1.7 Direct identification of miRNA:target interactions

CLIP methods can be adapted to capture miRNA:target duplexes via an addition of ligation step. RNA ligase connects 3'end of a miRNA and 5'end of a bound target RNA, generating a chimeric read. Thus, a chimeric read (in other words chimera) is composed of two parts with different genomic origins: miRNA part and target part. Further, these miRNA:target chimeras can be computationally identified. As a result, some of the Ago binding sites are directly assigned to a cognate miRNA. The first ligation-based identification of miRNA:target interactions was performed in David Tollervey lab (Helwak et al., 2013). The authors reported around 18,000 miRNA:mRNA pairs in HEK 293 cells. Inspired by their results we set out to explore miRNA targetome in living animal *C. elegans*. We modified *in vivo* PAR-CLIP, previously established in our lab (Jungkamp et al., 2011) with ligation step. I also designed, developed and tested the computational tool called 'ChiFlex' to extract miRNA:target chimeras from sequencing pool with controlled false discovery rate. Our efforts resulted in around 3 600 individual miRNA:target interactions supported by more than 5 000 chimeric reads in *C. elegans*. Surprisingly, comparable numbers of miRNA:target chimeras were discovered in both genuine (with ligation step) and control (without ligation step) experiments. We hypothesized, that there should be an active endogenous ligase responsible for the production of chimeras. In line with this hypothesis we re-analyzed previously published Ago-CLIP datasets in order to find yet hidden miRNA:target chimeras. Indeed, we discovered thousands of the interactions in various biological systems (human brain, human cell lines, mouse cell lines, *C. elegans*, virus-infected cells). Further we proved that our findings represent true endogenous binding events. (1) The discovered interactions follow the established rules of miRNA targeting: the majority of them utilize seed region as recognition motif and the binding sites tend to reside on 3'UTRs of protein coding genes. (2) The interactions are specifically conserved and enriched in previously validated miRNA:mRNA pairs. (3) The expression of genes associated with a particular miRNA is significantly shifted in the miRNA perturbation experiments. (4) We also validated several miRNA:mRNA interactions using luciferase reporter assay. Thus, endogenous miRNA:target interactions can be computationally recovered from conventional Ago-CLIP experiments with high specificity.

High specificity of our findings, as well as direct assignment of miRNA identity to the binding sites, allowed us to explore some aspects of miRNA biology, which were unachievable beforehand. For example, we tested the distribution of mismatches along the seed and cooperativity between miRNA family members. We also analyzed miRNA targeting in two systems with compelling post-transcriptional regulation: virus-infected cells and mammalian brains. For the former we characterized a viral miRNA mimicking the host one, and a miRNA with a shifted seed region. For the latter we

recovered a complex regulatory interplay involving two miRNAs and two non-coding RNAs. Thus, chimera-based method is a powerful tool to explore miRNA interactions, including those, critically important for disease and development.

2. Materials and methods

2.1 Ligation iPAR-CLIP

Worm culturing, labeling, crosslinking, homogenization

C. elegans transgenic animals expressing GFP::ALG-1 fusion proteins were used for experiments producing standard iPAR-CLIP and ligation iPAR-CLIP and control samples. Per sample 150,000 synchronized L1 worms were grown in liquid culture (S-Basal: 100 mM NaCl, 6 mM K_2HPO_4, 44 mM KH_2PO_4, 5 mg/L cholesterol supplemented with 3 mM $MgCl_2$, 3mM $CaCl_2$ and 10 mM K-Citrate (pH 6)) containing 3 mM 4SU (4-thiouridine) on a rotary shaker at 200 rpm and 20ºC until L3 stage. Liquid cultures usually contained 3,000 worms per ml and 1 ml *E. coli* OP50 (OD600 2.3) per 1,000 worms. L3 staged worms were transferred to NGM (nematode growth medium agar) plates and crosslinked on ice using a Stratalinker (Stratagene) with customized 365 nm UV-lamps (energy setting: 3 J/cm^2). Worms were lysed on ice by douncing in NP40 lysis buffer for 15 minutes (final: 50 mM HEPES-K pH 7.5, 150 mM KCl, 1 mM MgCl2, no EDTA, 0.5% (v/v) NP-40, 0.5 mM DTT, protease inhibitor cocktail (Roche)). Per volume settled worms (in 150 mM NaCl) 1 volume 2x NP-40 lysis buffer was used. Lysates were cleared by 10 minutes centrifugation with 13,000 rpm at 4ºC.

RNase treatments and immunoprecipitation

Cleared lysates were treated with RNase T1 (Fermentas) (final concentration 1 U/µl) for 15 min at 22ºC. GFP::ALG1 fusion proteins were immunoprecipitated for 2h at 4ºC using anti-GFP antibody (Roche, Cat. No. 11814460001) coupled to Protein G magnetic beads (Invitrogen). For each sample (1 ml cleared lysate obtained from 150,000 worms), 50 µl beads and 25 µg antibody were used. Immunoprecipitates were washed 3x with cold 1 ml IP-buffer (50 mM HEPES pH 7.5, 300 mM KCl, 0.05% (v/v) NP40 substitute, 0.5 mM DTT, protease inhibitor cocktail (Roche)). After a second treatment with RNase T1 (20 U/µl) for exactly 12 min at 22ºC, beads were washed 8x with cold High-Salt buffer (50 mM HEPES pH 7.5, 500 mM KCl, 0.05% (v/v) NP40 substitute, 0.5 mM DTT, protease inhibitor cocktail (Roche)), 3x with 1 ml cold NEB3 buffer (100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM DTT, pH 7.9) and then resuspended in 1 original bead volume cold NEB3 buffer.

Ligation of complete miRNAs to bound target sites

The 5'end of target RNAs are supposed to ligate to the 3'hydroxylated end of full-length miRNAs. The 5'ends of the miRNAs are incorporated in the MID domain of the protein and thereby inaccessible for28 ligation. To prevent circularization of ALG-1 bound target RNAs during the treatment with the T4 RNA ligase, a CIP treatment (part of the original iPAR-CLIP/PAR-CLIP protocol) was skipped and immunoprecipitates were treated with T4 PNK phosphatase minus (NEB, M0236) for 40 minutes at 10ºC (1 U/ul in 1 original bead volume NEB3 buffer) before ligation. As a result, their 3'ends remain inaccessible (2'3'cyclic phosphates or 3'phosphates) for the T4 RNA ligase, while phosphorylation of hydroxylated 5'ends prepares them for intermolecular ligation via the T4 RNA ligase. Phosphorylation was carried out for 40 min at 10ºC in NEB3 buffer (1 original bead volume) containing 0.5 mCi/ml γ^{32}P-ATP. Subsequently, nonradioactive ATP was added to 100 uM and the incubation was continued for 10 min at 10ºC. Beads carrying immunoprecipitated ALG-1 RNA complexes were washed 5x with 1 ml cold NEB3 buffer and resuspended in 10 original bead volumes ligation reaction buffer (for a 500 ul reaction: 50 ul T4 RNA ligase buffer (10x, Thermo Scientific), 60 ul PEG 8000 (50%), 5 ul KCL (1M), 12.5 ul RNasin Plus (40U/ul), 322.5 ul H2O).

For the ligation iPAR-CLIP sample, T4 RNA ligase (Thermo Scientific, EL0021) was added to 1 U/ul (50 ul for a 500 ul reaction), while the control sample was supplemented with the same volume 50 mM KCl (equivalent to the salt concentration of the enzyme storage buffer). The samples were incubated at 5ºC for 14 hours at 7 rpm on a rotating wheel. Thereafter beads were washed 5x with 1 ml cold NEB3 buffer, resuspended in 1 original bead volume NEB3 buffer and subjected to 3'end dephosphorylation with T4 PNK (NEB, M0201) for 40 minutes at 10ºC without ATP, to prepare target RNA 3'ends for addition of adapters (conversion of 2'3'cyclic phosphates/3'phosphates to 3'OH). Samples were washed 3 times with 1 ml cold NEB3 buffer and resuspended in 50 ul NuPAGE LDS Sample Buffer (Invitrogen, Cat. no. NP0007).

Denaturing protein purification, RNA isolation, cDNA library preparation, and PCR amplification (20 cycles) were performed as described previously (Hafner et al., 2010) with the difference of excising products with an RNA insert size of 20-35 nts and 35-60 nts (optionally combined in a 1:2 ratio for sequencing; majority of chimeras found in the later fraction). Libraries were sequenced on a Genome Analyzer II (Illumina) with 100 cycles.

2.2 Computational pipeline

Read processing

cDNA libraries produced from 20-60 nts long RNA fragments were Solexa sequenced with 100 cycles. Raw sequencing data from various CLIP studies were downloaded from GEO (GSE28865 Kishore et al., 2011; GSE41437 Skalsky et al., 2012; GSE32113 Gottwein et al., 2011; GSE41288 Loeb et al., 2012; GSE43574 Memczak et al., 2013; http://icb.med.cornell.edu/faculty/betel/lab/Data.html for data from Lipchina et al., 2011; http://ago.rockefeller.edu/rawdata.php for data from Chi et al., 2009). 3'adapter sequences and reads shorter than 15 nucleotides were removed using Flexbar (Dodt et al., 2012) and sequence reads were sorted according to their barcodes. Adapter concatemers were detected by scanning for overrepresented nucleotide stretches and removed. Reads of identical sequence were collapsed. For sequencing libraries generated from standard iPAR-CLIP and iPAR-CLIP ligation and control samples adapters ending with two random nucleotides were used, which allowed collapsing reads that were derived from individual PCR templates independent of potential PCR overamplification.

Detection of complete and truncated miRNAs in sequencing reads

For each mature miRNA annotated in miRBase version 19 for human, mouse and *C. elegans*, we generated "anchors", defined as all possible 12 nts windows from miRNA sequences. Thus, one anchor might relate to different miRNAs and the search algorithm references to all of them when presence of the anchor is reported. Collapsed sequencing reads and control reads (generated from sequences of collapsed reads by permuting dinucleotides) were searched for anchors. Reads containing anchors were locally aligned (Smith-Waterman: match bonus 2, mismatch penalty 5, gap open 6, gap extension 4) to the miRNAs they referenced to. Only the miRNA with the best alignment was considered to be inside the read. The part of the read that could be aligned to the miRNA was termed "miRNA match part". All reads having the same alignment score were grouped and for each group a FDR was calculated by dividing the number of mapped control reads by the number of mapped real sequence reads. Groups with a FDR<0.05, were considered "reliable", mainly comprising reads aligned with a continuous stretch of at least 14 nts or with 17 nts interrupted by one mismatch. They were further characterized in terms of: position of miRNA match part within the read and position of the miRNA match part in the complete miRNA sequence. Hereby determined features were considered as hallmarks if present in more than 1% of reliable reads. Usually features found to be enriched were: the miRNA being at the 5'end of the sequencing read and the 1st nucleotide of miRNA being present. Then, for every alignment score group, reads with a FDR>=0.05 were reexamined for the presence of these hallmarks. An individual FDR was calculated for the subset of reads found to

possess the respective hallmarks and they were included in further analysis if their FDR<0.05. For subsequent analysis, the sequence parts downstream the miRNA match parts were considered as "target candidates", if their length >=15 nts. The actual boundary between miRNA and target parts inside the chimeric reads was assigned after mapping target candidates (for details see 2.4).

Identification of target RNAs ligated to miRNAs

Target candidates found via the search for miRNA parts within sequencing reads (see 2.1) were mapped to Argonaute clusters from all studies cited in Table 1, plus our own data from *C. elegans*. The last nucleotides of the miRNA match part could in theory, also belong to the downstream target candidate sequence. For target candidates from datasets that were produced using RNase T1 in the CLIP experiment, we took advantage of the nucleotide bias known for this enzyme. Because RNase T1 preferentially cuts after guanosines, we considered the last nucleotide of the miRNA match part to belong to the target candidate sequence, if the preceding nucleotide in the miRNA sequence was a G. This increased the chance for mapping, especially for short target candidate sequences. Sequencing reads of PAR-CLIP samples contain characteristic T to C mutations caused by crosslinking of photoreactive nucleoside 4-thiouridine. To use this feature and to facilitate mapping, target candidate sequences from PAR-CLIP datasets were mapped using variations of their sequence, each containing one "inverse" C to T mutation.

Mapping was performed using Bowtie2 in the local alignment mode. Bowtie2 parameters were chosen to be soft (window length 13, window step 1, alignment score 30) to enhance sensitivity. False positive mappings were excluded only at a later step of analysis. Usually these settings translated into the acceptance of >=15 nts perfect match or >=19 nts match containing one 1nt mismatch, deletion or insertion. Only uniquely mapped target candidate sequences were considered further.

To estimate a false discovery rate and to filter out unreliable mappings, control sequences (generated by permuting dinucleotides of target candidate sequences) were mapped and their alignment score was used to define reliable mapping. In detail, the cutoff was chosen to be the lowest alignment score that still guaranteed an FDR<0.05 (number of mapped sequences with alignment score >= cutoff divided by number of mapped control reads with alignment score >= cutoff). More than 90% of reliable mappings had no unmappable nucleotides between miRNA match part and target sequence. Therefore we used this feature combined with the minimum alignment score, which still guarantees a FDR<0.05 to rescue reads that did not pass the original cutoff. In addition, sequence reads that did not pass the cutoff but mapped to the same genomic loci as reliably mapped sequences, were also included in further analysis, in order to correctly assess the number of reads supporting an interaction.

Target sequences that were ligated to the same miRNA and mapped to the same genomic loci were

gathered (clustered). They were extended to include nucleotide positions that might have been originally paired to the miRNA but had been cut away by RNases or had been positioned beyond sequencing length. To investigate how far target sequences should be extended downstream, we elongated them stepwise (one nt position at a time), always inspecting the gain of seed matches compared to a control in which nucleotides were added randomly. We found that downstream extension of 0-8 nts led to a significant increase in seed matches. To estimate by how many nucleotides the recovered target sequences should be elongated at the 5'end, we analyzed seed-match containing target sequences. An upstream extension of 8-12 nts enabled 90% of seed-match containing target sequences to possess at least 24 nucleotides upstream the 3'most position of the seed-match. Consequently sequences were extended upstream by 8-12 nts and were further referred to as "target site", the pair of miRNA and its target site was termed "interaction". Pre-miRNAs, which usually constituted less that 2% of interactions, were also excluded from downstream analysis. A more focused search procedure was performed for interactions between miRNAs cel-let-7, cel-lin-4 and *C. elegans* transcripts lin14, lin-28, lin-41. We run the same analysis as described here, but with search spaces limited to these miRNAs and transcripts. This enables the relaxations of cutoffs and enabled the recovery of additional miRNA-interactions.

Identification of miRNA:target boundaries inside chimeric reads

When searching for miRNA parts inside sequencing reads, miRNAs were aligned as far as possible into 3' direction of reads. Occasionally this can lead to the misannotation of the miRNA identity (especially between miRNA family members) when nucleotides belonging to the ligated target sequence are considered to be nucleotides of the miRNA. Therefore, the boundary between miRNAs and target sites were reassigned after the mapping of target candidates. For reads in which nucleotides cannot be assigned unambiguously to the miRNA or the target, we applied the following heuristics. For reads generated in experiments using RNase T1 we assigned all nucleotides downstream the first guanosine within the stretch of ambiguous nucleotides, to be part of the target. In reads from experiments using RNase A, all ambiguous nucleotides were assigned to the target part in order to prevent misannotation of miRNA identities. A few chimeras were discarded as the newly assigned miRNA sequence were too short to pass the requirements for miRNA identification.

2.3 Downstream analysis of ligation products

RNase produced cleavage sites involved in formation of chimeras

To investigate whether RNase-produced cleavage sites are involved in the formation of ligation products we determined frequencies of all four nucleotide types at the end of miRNA match parts and 5'upstream the target sequences and quantified by how many nucleotides the miRNAs recovered in ligation products were truncated. For this analysis data from the iPAR-CLIP control sample was used and chimeras in which the last nucleotide of the miRNA match part could also constitute the first nucleotide of the target sequence, were excluded.

Binding free energy and hybridization pattern

For each interaction binding free energy and hybridization pattern were predicted by RNA hybrid (version 2.1.1), allowing G:U pairing. For calculation of binding free energy the first nucleotide of the miRNA was excluded as structural studies suggest, that it is not involved in base pairing with the target (Elkayam et al., 2012). Shuffled sequences (dinucleotides in target sequences are permuted) and shuffled interactions (targets are swapped between miRNAs) served as controls and multiple rounds of generation and resampling of control interactions served to find a consensus control binding free energy distribution and control hybridization pattern. The distribution of binding free energies was smoothed. For hybridization profiles summarized over all miRNA interactions, the predicted frequency of a miRNA position being bound is plotted along the miRNA length. For instance, if nucleotide position 3 of the miRNA is bound in an interaction, the value in bin "3" is incremented. For clarity, the hybridization pattern of control interactions was subtracted from the pattern of interactions derived from chimeras.

Detection of miRNA complementarities

Interactions were checked for presence of complementarities to the seed. Specifically, for each interaction the target sequence was screened for reverse complementary matches to miRNA nucleotides: 2-7, 2-7 with 1 nt mismatch in any position, 2-7 with 1 nt bulge in the target, 2-8 with 2 mismatched nucleotides. Again, shuffled target sequences (dinucleotides in target sequences are permuted) and shuffled interactions (targets were swapped between miRNAs) served as controls.

miRNA interactions recovered from CLIP data by Kishore et al., 2011 were analyzed for stretches of complementarity between the 3'part of the miRNA (starting from miRNA nt position 9) and the target site (for interactions with a perfect seed match or a seed match containing 1 nt mismatch, only the target sequence upstream the match was considered). Per miRNA consecutive windows of 4 nts were

generated. If a match to the particular window (for example 11-14 nt in miRNA) is found in the corresponding target, all values corresponding to that window (11-14 positions) are incremented. Shuffling target sequences (dinucleotides were permuted) served as control. Per interaction 100 control interactions were generated and per miRNA position the maximum value was chosen as a control signal. A miRNA position was reported to have significant complementarity, if its value is higher than the control value. Only significant complementarities of at least 4 nts originating from the analysis of at least 20 interactions were reported.

Sites targeted by miRNAs of the same seed family

miRNA families are defined as the set of miRNAs that share the same seed sequence (nts 2-7). Per miRNA family we determined the number of instances in which the same genomic locus is targeted by at least two miRNA family members. This analysis was performed for chimera-derived *C. elegans* miRNA interactions and for interactions recovered from human AGO PAR-CLIP and HITS-CLIP data (Kishore et al., 2011). Results were depicted for the five most abundant miRNA families recovered in chimeras. Multiple rounds of generation and resampling of control interactions served to assess statistical significance.

Quantification of T to C conversions relative to the seed match

For chimeras containing prefect seed (2-7) complementarity, the distance between T to C conversion and the position that is complementary to miRNA position 2 (3'most nucleotide of the seed match) was calculated. Summarized per nucleotide position in the target and normalized to the number of chimeras, we received local enrichment values for T to C conversions. These were further normalized to the positional frequencies of thymidines in the target sequences. An equivalent analysis was performed for interactions having a 2-7 match with 1 nt mismatch at any position of the seed.

Positional mismatch frequency within the seed

For interactions with seed matches (2-7) containing one mismatched nucleotide, the position of the mismatch was analyzed. To rule out potential bias introduced by adapter ligation, only interactions supported by at least one chimeric read containing more than 3 nts downstream of the seed match were considered. The mismatch frequencies per position (2-7) were calculated per miRNA family (defined by their common hexamer seed sequence) and averaged across families. Thereby, each miRNA family contributed equally, eliminating biases due to miRNA abundance. This analysis was performed on miRNA interactions recovered from human AGO PAR/HITS-CLIP data (Kishore et al.,

2011), mouse AGO HITS-CLIP data (Chi et al., 2009, Loeb et al., 2012) and *C. elegans* ALG-1 iPAR-CLIP data of the present study. For the smaller number of mouse miRNA-chimeras, we dropped the requirement of having at least 3 nts downstream the seed match. Including this requirement decreased the number of chimeric reads used for the analysis but did not alter the result.

Conservation analysis of miRNA interactions

Conservation analysis was performed for chimera-derived *C. elegans* miRNA interactions and for interactions recovered from human AGO PAR-CLIP and HITS-CLIP data (Kishore et al., 2011). Only interactions residing in 3'UTRs (Gerstein et al., 2010 for *C. elegans* and hg19 refGene table based on Pruitt et al., 2005 downloaded for human 3'UTRs from UCSC table browser at ucsc.edu, respectively) were considered to circumvent conservation originating from coding sequences and from cis-regulatory elements of translation initiation. To investigate the conservation of perfect hexamer seed matches from *C. elegans* and human miRNA interactions, 4 nematode species (*C. brenneri, C. briggsae, C. japonica, C. remanei*) and 31 vertebrate species (tarSyr1, micMur1, tupBel1, mm9, rn4, dipOrd1, cavPor3, speTri1, oryCun2, ochPri2, vicPac1, turTru1, bosTau4, equCab2, felCat3, canFam2, myoLuc1, pteVam1, eriEur1, sorAra1, loxAfr3, proCap1, echTel1, dasNov2, choHof1, macEug1, monDom5, ornAna1, galGal3, taeGut1, anoCar1) were checked for the seed sequence at the same position in multi-species alignment of 3'UTRs (data taken from ce6 6way multiz nematode and hg19 46way multiz vertebrate alignment blocks tables at ucsc.edu, Karolchik et al., 2014; stitched via internal_maf_to_merged_fasta.py script from the Galaxy pipeline), respectively. To investigate the conservation of seed matches with 1 nt mismatch (at any position in 2-7), aligned hexamers were checked for the presence of the identical 1mm seed match or a perfect 2-7 seed match. Conservation of predicted seed matches (perfect or with 1 mismatch at the same position and of the same type as in the chimera-derived interaction, respectively) in 3'UTRs and AGO binding sites was used to evaluate the increment in conservation gained when using miRNA targeting information from our analysis of miRNA chimeras.

2.4 Mutagenesis and reporter assays to test miRNA interactions

Construction of 3'UTR reporter vectors

WT and miRNA binding site mutant 3'UTR sequences were amplified from BC-1 genomic DNA using the primers specified below and inserted between the XhoI and NotI sites of the previously described firefly luciferase reporter vector pLSG (Gottwein and Cullen, 2010; Gottwein et al., 2007). Resulting constructs were confirmed by Sanger sequencing.

miR-K11 binding sites:

Reporter	outer primers	inner mutant primers
BCL2	F: AGAGACTCGAGGTGTGGCCTTGGCCCACCTG R: TCTCTGCGGCCGCTGCGGAATTGCCCAGGGACG	N/A
BCL2 mutant	F: AGAGACTCGAGGTGTGGCCTTGGCCCACCTG R: TCTCTGCGGCCGCTGCGGAATTGCCCAGGGACG	F: ATTGATGGAATAACTCTGTGCGTTATTTCGTATATATACCATTTATCTGTAT R: ATACAGATAAATGGTATATATACGAAATAACGCACAGAGTTATTCCATCAAT
CLCN3	F: GAGAGGCGGCCGCTGGAGGAGTTGTTTGGGGAGGGA R: CTCTCGAATTCTCGGTTTTGAGCCACACGGCA	N/A
CLCN3 mutant	F: GAGAGGCGGCCGCTGGAGGAGTTGTTTGGGGAGGGA R: CTCTCGAATTCTCGGTTTTGAGCCACACGGCA	F: TTTAATTCATGAATTGTATACTTAACGTAATCCTTTCTACATTCCAGAAG R: CTTCTGGAATGTAGAAAGGATTACGTTAAGTATACAATTCATGAATTAAA
KHDRBS1	F: AGAGACTCGAGCATGAGGGGAAAATATCAG R: TCTCTGCGGCCGCGACAAACTGCTGCAGACAT	N/A
KHDRBS1 mutant	F: AGAGACTCGAGCATGAGGGGAAAATATCAG R: TCTCTGCGGCCGCGACAAACTGCTGCAGACAT	F: GATTTCTTGTATCTCCCAACTTTGCTCTACGTAATCCCACAACAGACAAGTAA R: TTACTTGTCTGTTGTGGGATTACGTAGAGCAAAGTTGGGAGATACAAGAAATC
MYB	F: AGAGACTCGAGGACATTTCCAGAAAAGCATT R: TCTCTGCGGCCGCGCTACAAGGCAGTAAGTAC	N/A
MYB mutant	F: AGAGACTCGAGGACATTTCCAGAAAAGCATT R: TCTCTGCGGCCGCGCTACAAGGCAGTAAGTAC	F: CATTTTATGAGTTTTCTGTTACCCTTTTAAAAAATAACTTACTGTAAGAAATAGTTTTAT R: ATAAAACTATTTCTTACAGTAAGTTATTTTTTAAAAGGGTAACAGAAAACTCATAAAATG
RORA	F: AGACTCGAGTATGATTTCCATTATGCC R: TTCGCGGCCGCGAGTCCATATTTAACTAC	N/A
RORA mutant	F: AGACTCGAGTATGATTTCCATTATGCC R: TTCGCGGCCGCGAGTCCATATTTAACTAC	F: CATAGCAGTAGCAACAATAGGATAATAATATATTACAGGGTAAA R: TTTACCCTGTAATATATTATTATCCTATTGTTGCTACTGCTATG

STK38L	F: AGAGACTCGAGTGCCTGTGTGTGCTGTGGCT R: TCTCTGCGGCCGCCACAACCCCCTGGCCTGCAT	N/A
STK38L mutant	F: AGAGACTCGAGTGCCTGTGTGTGCTGTGGCT R: TCTCTGCGGCCGCCACAACCCCCTGGCCTGCAT	F: GAACTTCTTTTTTTAACAAGACCAGATGCGATTATTTTAATTTGATTATGG R: CCATAATCAAATTAAAATAATCGCATCTGGTCTTGTTAAAAAAAGAAGTTC
ZNF330	F: AGAGACTCGAGGGGAGCTGCTCTGGTGGCCG R: TCTCTGCGGCCGCTGTCCTGGGTAAAGCTTCTG	N/A
ZNF330 mutant	F: AGAGACTCGAGGGGAGCTGCTCTGGTGGCCG R: TCTCTGCGGCCGCTGTCCTGGGTAAAGCTTCTG	F: GTAGCGTTTTTATAGAACTCCTAATCACCGATATGCGATAAGAAAAATGAGTTTC R: GAAACTCATTTTTCTTATCGCATATCGGTGATTAGGAGTTCTATAAAAACGCTAC
PUM2	F: AGAGACTCGAGTGCTGGCAAAGCACAGAATGCCT R: TCTCTGCGGCCGCGCCAGGTCAAAGAGGGCAGGG	N/A
PUM2 mutant	F: AGAGACTCGAGTGCTGGCAAAGCACAGAATGCCT R: TCTCTGCGGCCGCGCCAGGTCAAAGAGGGCAGGG	F: GTGGTTTGAGATGAAAAGAACATGAAGTGATTTACAGTAGATGTGGTTTT R: AAAACCACATCTACTGTAAATCACTTCATGTTCTTTTCATCTCAAACCAC
YWHAZ	F: AGAGACTCGAGCTTTCTCTACAGCTTTTC R: TCTCTGCGGCCGCGAGCATTTACAGTGTACT	N/A
YWHAZ mutant	F: AGAGACTCGAGCTTTCTCTACAGCTTTTC R: TCTCTGCGGCCGCGAGCATTTACAGTGTACT	F: GTGTTCCATTTAAAATTTTGTGATATGAATGATTCTAACTTAGGAAGCCACA R: TGTGGCTTCCTAAGTTAGAATCATTCATATCACAAAATTTTAAATGGAACAC

miR-K4-3p binding sites:

Reporter	outer primers	inner mutant primers
TRIM33	F: AGAGACTCGAGTTCCAGAAAACACTTCCTCAC R: TCTCTGCGGCCGCGGGGTGGTAAAGGTGGGTTT	N/A
TRIM33 mutant	F: AGAGACTCGAGTTCCAGAAAACACTTCCTCAC R: TCTCTGCGGCCGCGGGGTGGTAAAGGTGGGTTT	F: CATTTGCTGAGCCTGTTTTTTACATGAATGTACCGAATTACTAATGT TCCTATCAAGAA R: TTCTTGATAGGAACATTAGTAATTCGGTACATTCATGTAAAAAACAGGCTCAGCAAATG

KBTBD11	F: AGAGACTCGAGAAGGATGT TAACTGATGTAG R: TCTCTGCGGCCGCGTTGGAT TCGGAGTTCAAAT	N/A
KBTBD11 mutant	F: AGAGACTCGAGAAGGATGT TAACTGATGTAG R: TCTCTGCGGCCGCGTTGGAT TCGGAGTTCAAAT	F: GGGAGATTAATTTTTACAATTTCTGAAGAAAGTGTGGCC GAGTGACAGTG R: CACTGTCACTCGGCCACACTTTCTTCAGAAATTGTAAAAA TTAATCTCCC
N/A		
c14orf2 mutant	F: AGAGACTCGAGCCAGATTT ACTTGGAGTACA R: TCTCTGCGGCCGCTCTCTTA TTTATCCTCATAA	F: GGACTTGGTGATCAGGATCCACATATCCACTTGACTAAT ACTGCTCAATAAACGTTTA R: TAAACGTTTATTGAGCAGTATTAGTCAAGTGGATATGTG GATCCTGATCACCAAGTCC

miR-K1 binding sites:

Reporter	**outer primers**	**inner mutant primers**
HDDC2	F: AGAGACTCGAGGACACTCTCTAAATTGC R: TCTCTGCGGCCGCGGTCTCAAGATTTATAA GTC	N/A
HDDC2 mutant	F: AGAGACTCGAGGACACTCTCTAAATTAGTG CACGGAAACTTCAAACATTATTTTTCCATTT C R: TCTCTGCGGCCGCGGTCTCAAGATTTATAA GTC	N/A
PIP5K1C	F: AGAGACTCGAGTGTTGTCTCCAAGGCCCTT T R: TCTCTGCGGCCGCCGAGGTTTGCAGTTTCA TTT	N/A
PIP5K1C mutant	F: AGAGACTCGAGTGTTGTCTCCAAGGCCCTT T R: TCTCTGCGGCCGCCGAGGTTTGCAGTTTCA TTT	F: AGAACGTGGGCAGTGTCCAAAGCGATTTAT CGACGAGACTAAAAGGCGTTTGCTCT R: AGAGCAAACGCCTTTTAGTCTCGTCGATAA ATCGCTTTGGACACTGCCCACGTTCT
SNRK	F: AGAGACTCGAGGAACTATAGGGCCTAGTA CA R: TCTCTGCGGCCGCTACTTGAAGAAACAGTG ACC	N/A
SNRK mutant	F: AGAGACTCGAGGAACTATAGGGCCTAGTA CA R: TCTCTGCGGCCGCTACTTGAAGAAACAGTG ACC	F: TAATCAAAGAACTCTTGCTTTAAATATGAT GAGATCAAAGACTGTTTTTGACCAGATA R: TATCTGGTCAAAAACAGTCTTTGATCTCAT CATATTTAAAGCAAGAGTTCTTTGATTA

* Mutant primer 1866 repaired a SNP that occurred in the BC-1 cell line, resulting in an additional 1 nt difference between WT and mutant vector just downstream of the miR-K1 binding site.

miR-K3 binding sites:

Reporter	outer primers	inner mutant primers
IRF4	F: GAGAGCTCGAGCTTACA GTATTGTTACCACC R: TCTCTGCGGCCGCCTCT CTGAAGGAACGTAACC *2	N/A
IRF4 mutant	F: GAGAGCTCGAGCTTACA GTATTGTTACCACC R: TCTCTGCGGCCGCCTCT CTGAAGGAACGTAACC *2	F: CAAACTGAGGTAGATAATGCTATACTATCATTGGTATACAC CATAAATTTTTATGTAAATTGCTCTGC R: GCAGAGCAATTTACATAAAAATTTATGGTGTATACCAATGA TAGTATAGCATTATCTACCTCAGTTTG
TMEM173	F: GAGAGCTCGAGGACCCA GGGTCACCAGGCCA R: TCTCTGCGGCCGCGGAA ACCGCAAGTGAGAGGG	N/A
TMEM173 mutant	F: GAGAGCTCGAGGACCCA GGGTCACCAGGCCA R: TCTCTGCGGCCGCGGAA ACCGCAAGTGAGAGGG	F: GTGAAATGGGATCATAATCACATAATTACCAGACTTACGCT ATTAGTGAGGACTGAGTGTGTGGAAG R: CTTCCACACACTCAGTCCTCACTAATAGCGTAAGTCTGGTA ATTATGTGATTATGATCCCATTTCAC
TRIB1	F: AGAGACTCGAGTCCCCA AAACCTCAGAAACCTC R: TCTCTGCGGCCGCCTAG ACTGTAAGTCCTGAGGC C	N/A
TRIB1 mutant	F: AGAGACTCGAGTCCCCA AAACCTCAGAAACCTC R: TCTCTGCGGCCGCCTAG ACTGTAAGTCCTGAGGC C	F: GAATGATTATTGGCAATATTATATTGAAAATAACATGGGAC TTTGAGAAGAGGG R: CCCTCTTCTCAAAGTCCCATGTTATTTTCAATATAATATTGC CAATAATCATTC
USP33	F: AGAGATCTAGATTTTTA GGATGTAGAGAGTTC R: TCTCTGCGGCCGCGAGC TAATTTTAACAACATTG	N/A
USP33 mutant	F: AGAGATCTAGATTTTTA GGATGTAGAGAGTTC R: TCTCTGCGGCCGCGAGC TAATTTTAACAACATTG	F: GCTTATTAAAATTTACAAAATTAATTTTTTAGTAAATCAAG CATATATTTAGTTTGAGTGGAA R: TTCCACTCAAACTAAATATATGCTTGATTTACTAAAAAATT AATTTTGTAAATTTTAATAAGC
AK3	F: AGAGACTCGAGGGAGA AATGTGTGTAACTATT R: TCTCTGCGGCCGCCACT GCTCACTTTGATTTCC	N/A

AK3 mutant	F: AGAGACTCGAGGGAGA AATGTGTGTAACTATT R: TCTCTGCGGCCGCCACT GCTCACTTTGATTTCC	F: CATTCTTTATAAAACTTTCTATAAATAAATATTTAGTATTTA ATCTTATGTGCTTTCTAAAAA R: TTTTTAGAAAGCACATAAGATTAAATACTAAATATTTATTT ATAGAAAGTTTTATAAAGAATG

*[2] Due to the proximity of the binding site to the polyA signal, this construct also includes 101 bp of genomic sequence downstream of the annotated polyadenylation signal to achieve authentic polyadenylation.

Reporter assays

Reporter assays were performed in 293T cells, essentially as described previously (Gottwein and Cullen, 2010). Briefly, unmodified pLSG firefly luciferase vector, WT or binding site mutant 3'UTR reporter vectors were co-transfected with the internal Renilla luciferase control vector pLSR and mirVana™ miRNA mimics (Life Technologies) or pLCE-based miRNA expression vectors. For miRNA mimics, each well of a 24 well plate was co-transfected with 5 pmoles mimic, 0.3 ug irrelevant plasmid DNA, 2.5 ng pLSG vector and 5 ng pLSR vector using 1 ul Lipofectamine 2000 (Life Technologies). For miRNA expression vectors, each well of a 24 well plate was co-transfected with 0.4 ug empty miRNA expression vector pLCE, WT or seed mutant miR-K11 expression vectors, 2.5 ng pLSG vector and 5 ng pLSR with 1 ul Fugene6 (Roche). miR-30-embedded WT and seed mutant miR-K11 expression vectors were described and validated previously (Gottwein et al., 2007; 2011). In seed mutant miR-K11, nts 2-7 were changed from UAAUGC to UAUUCC. Dual luciferase assays were performed 48 hours after transfection using the Dual-Luciferase® Reporter Assay System (Promega) as instructed. Resulting firelfy reporter activities were first normalized to those from the internal control vector pLSR. Resulting values for miRNA co-expression were further normalized to those from samples that received control mimic or empty miRNA expression vector pLCE. Finally, values for WT 3'UTR reporters were normalized to those obtained for the corresponding miRNA binding site mutant vector, to isolate the regulatory potential of each specific chimera-identified miRNA binding site.

2.5 ChiFlex methods

Here the methods used for the second part of Results section are disclosed.

Preprocessing

Preprocessing is not included in ChiFlex package. However, it is important to purify the sequencing data prior submitting them to ChiFlex. The more barcodes, primers and adapters are in sequencing reads the less is sensitivity of chimeras' detection. To purify the analyzed datasets we applied a combination of FastQC and Flexbar. The former detects barcodes, adapters and other overrepresented sequences, the latter excises these sequences out of the reads. For several datasets multiple rounds of adapter/barcode removal were applied, since AGO-CLIP datasets are often enriched in adapter/barcode concatemers. We tried to purify reads until at least 70% are exclusively composed of endogenous RNA insert.

Mapping to the miRNA sequences

ChiFlex utilizes Bowtie2 mapper in a local alignment mode. That is we look for valid alignments of the parts of the read, not the whole read. The first round of mapping is performed to miRNAs' sequences and miRNAs' sequences with introduced mutations ('true' and 'decoy' references). The second round of mapping is performed to a potential targetome (genome, transcripts, 3'UTRs, etc.) and to its 'decoy' counterpart. As a default ChiFlex calls Bowtie2 with the following settings for miRNA part detection:

bowtie2 -D 40 -L 12 -N 0 -S [path to the output] -R 4 -U [path to the reads] --min-score C,26 --no-unal -i C,1 -k 8 --norc --rdg 8,6 --rfg 14,8 --mp 5,5 -x [path to the Bowtie2 index] --local

Detailed description of Bowtie2 is provided via its manual page on Sourceforge.

Processing of the mapping hits

ChiFlex optionally applies a filter for repetitive sequences. For each aligned sequence it calculates Shanon entropy for nucleotide transition probabilities. As a default we require the entropy to be higher than 1.6. It is important that ChiFlex calculates the entropy for aligned sequences, not for the whole read. Otherwise the read partially composed of repetitive sequence can pass the filter and detection of this repetitive chunk as a part of chimera will be distorted.

ChiFlex has an option to keep non-unique mappings. It is especially valuable for miRNA part detection, since many miRNAs are arranged in families. Family members share significant sequence similarity with each other, and stretch of 13-17 nucleotides can be attributed to multiple miRNAs. ChiFlex does not remove such unambiguous mappings. Instead it collapses them to the unique mapping to one of the miRNAs, keeping information about the others in a separate file. Further this information is retrieved to reassign the number of unique reads supporting chimeras with particular miRNA. This strategy is useful not only for the mapping to miRNAs, but also for the mapping to a genome in a case of multi-copy genes, including pseudogenes.

Demultiplexing of the mapping hits

One sequencing read can be aligned to different loci on both 'true' and 'decoy' references, having multiple mapping hits. For each mapping hit ChiFlex assigns a score via the following formula: as*(4-log(qs+1)-log(rs+1)); as – alignment score, qs – start of the alignment on a read, rs – start of the alignment on a reference (miRNA). The mapping hits with the best scores (score > max(scores)-2) are selected. If multiple hits are selected, they are assigned as non-unique mappings. Otherwise selected hit is assigned to be 'true' or 'control' if it was mapped to 'true' or decoy reference respectively.

Filtering

ChiFlex employs Logic Rule Generator (LRG) to find the set of filters which guarantee false discovery rate (FDR) below a preset cutoff along with suboptimal sensitivity. As input LRG requires: 'true' objects, 'control' objects, list of objects' attributes which will be used for filtering. Attributes must be numerical (For example: 'Alignment Score', 'Position of the Alignment', 'Mapping quality'). Each object is converted into a tuple of its values for selected attributes. That is a mapping hit with an alignment score 32 and position of the alignment 2 is converted into (32, 2). These tuples can be considered as point coordinates in multidimensional space of attributes. Thus LRG creates a multidimensional grid, each cell of which is filled with the objects with the identical values of attributes.

LRG starts with the selection of the cell with the highest number of 'true' objects and false discovery rate ((number 'control' objects)/(number total objects)) lower than selected threshold (typically equal 0.05). This cell is the origin of the cluster. Further, this cluster is extended in one dimension. The extension is chosen in a way to capture as much as possible 'true' objects while keeping FDR below

the threshold. LRG gets an updated cluster and looks for the best extension for it. Then the cluster grows iteratively via one-dimensional extensions until the point when there is no possible extension (with FDR below the threshold) available. Finally LRG checks if the fraction of objects captured by the cluster is higher than 1%. If so, cluster boundaries are translated into the intersection of one-dimensional thresholds and all the objects from the cluster are removed from the grid. LRG has 200 trials to generate the rules. All valid rules (clusters) are united ('OR' logic) into a single rule (filter).

The following attributes are used to filter the mappings to miRNA reference: alignment score, position of the alignment on a miRNA, position of the alignment on a read. LRG provides the filter based on these attributes. Finally, ChiFlex applies the filter on mappings to the 'true' reference.

Target part detection

The reads harboring miRNA part are further mapped to a potential targetome. It can be a genome, a transcriptome, or a set of 3'UTRs. For the human datasets we used hg38 genome assembly and mm10 for the mouse dataset. As a default ChiFlex calls Bowtie2 with the following settings for target part detection:

bowtie2 --ignore-quals -D 35 -L 14 -N 0 -S [path to the output] -R 4 -U [path to the reads] --min-score C,32 --no-unal -i C,1 -k 6 -p 4 --rdg 8,6 --rfg 18,12 --mp 5,5 -x [path to the Bowtie2 index] –local

Since the size of the mapping reference may vary largely, '--min-score' can be adjusted respectively ('min-score' ~ $\log_4(L)$ | L is the length of the genome). However, according to our experience ChiFlex performs reasonably well with the default parameters. One have to remember that '--norc' option must be used to not map the reads to the reverse complement of the provided reference (any case, but genome).

Mapping hits are demultiplexed and filtered as described above. LRG is applied using the following attributes: alignment score, gap between miRNA and target part.

miRNA:target interactions recovery

Chimeric reads with identical miRNA and target are merged into single miRNA:target interactions. The number of the reads is assigned to the interaction as 'read support'. ChiFlex utilizes python wrapper (pybedtools) for bedtools merge command to find the overlapping regions.

Hybridization profile analysis

For each miRNA:target pair we ran RNAhybrid with the default parameters. We took miRNA:RNA duplex with the minimal free energy of binding and checked which nucleotides of miRNA are base-paired. Then we calculated the fraction of the interactions with the N^{th} nucleotide being base-paired, where N is a position of a nucleotide on miRNA sequence. To get the background profile we swapped the target sequences randomly between the miRNAs and calculated the pattern as described above. We repeated this procedure 100 times. Finally we averaged these 100 'shuffled' patterns into one background pattern.

Evaluation of miRNA binding modes usage

MiRNA:targets were checked to interact via the previously described binding modes: matches to the nucleotides 2-7, 2-8 and 2-9 with or without adenosine opposite to the first nucleotide on miRNA, match to the nucleotide 3-8 , match to the nucleotide 2-8 with one mismatch allowed. To get the background probabilities for the modes we swapped the target sequences randomly between the miRNAs and looked for the modes. We repeated this procedure 100 times. Finally, the probabilities of the modes were averaged across these 100 'shuffling' iterations.

MiRNA expression levels quantification

Sequencing reads from Ago-CLIP experiments are mapped to a reference composed of mature miRNA sequences and the corresponding decoy reference. Then, the mappings are selected and filtered in the way described above (ChiFlex, miRNA:target case). For the further quantification we select only those reads which have less than 14 unmapped nucleotides downstream miRNA alignment. That is, the reads which cannot be identified as chimeras. Then for the reads which were mapped to multiple miRNAs, we split the counts between these miRNAs proportional to their expression. Finally, we normalized the expression to the counts per million reads (CPM).

3. Results

3.1 ALG-1 iPAR-CLIP reveals thousands of miRNA binding sites in C. elegans

The main purpose of my Ph.D project was to computationally resolve and characterize miRNA interactions in various biological systems, including human, mouse, worm and virus-infected cells. I started with the analysis of the sequencing data coming from modified PAR-CLIP performed in a model organism *C. elegans*. The reason for choosing this system is transparency of these animals and ease of modified nucleosides delivery.

In vivo PAR-CLIP (iPAR-CLIP) in C. elegans has been already established in our lab (Jungkamp et al., 2011). Further, Stefanie Grosswendt modified it for Argonaute IP in a way to generate miRNA:target chimeras. The design of the experiment includes the following steps. Worms are kept in a medium containing 4-thiouridine (4SU) and incorporate these photoreactive nucleosides into newly transcribed RNAs. RNAs carrying 4SU are further crosslinked to the bound proteins by UV-irradiation. Worms are then homogenized and treated with mild concentration of RNase. The ALG-1 proteins (one of two C. elegans Argonautes) are immunoprecipitated with bound RNA pieces. T4 RNA ligase is added to link the 3'hydroxyl (3'OH) of miRNA and phosphorylated 5'end of its cognate target (Fig. 1A). Since the addition of the ligation step might lead to circularization of ALG-1 bound RNA, a phosphatase treatment is skipped to not create 3'OH ends on target RNA. Analogous enzyme is used instead (see methods) to phosphorylate the 5'ends (Fig. 1C). Ago:miRNA:target complexes are then subjected to the second round of RNase treatment. Finally, the RNA pieces bound to ALG-1 (including those ligated to miRNAs) are recovered under stringent conditions, went through cDNA amplification and finally sequenced (Fig. 1A,B).

Altogether three types of experiments were performed. Conventional iPAR-CLIP; modified iPAR-CLIP; modified iPAR-CLIP without ligation step. The first and the third experiments were considered as negative controls for the second in terms of miRNA:target chimeras discovery. All the protocols were exerted in two biological and technical replicates.

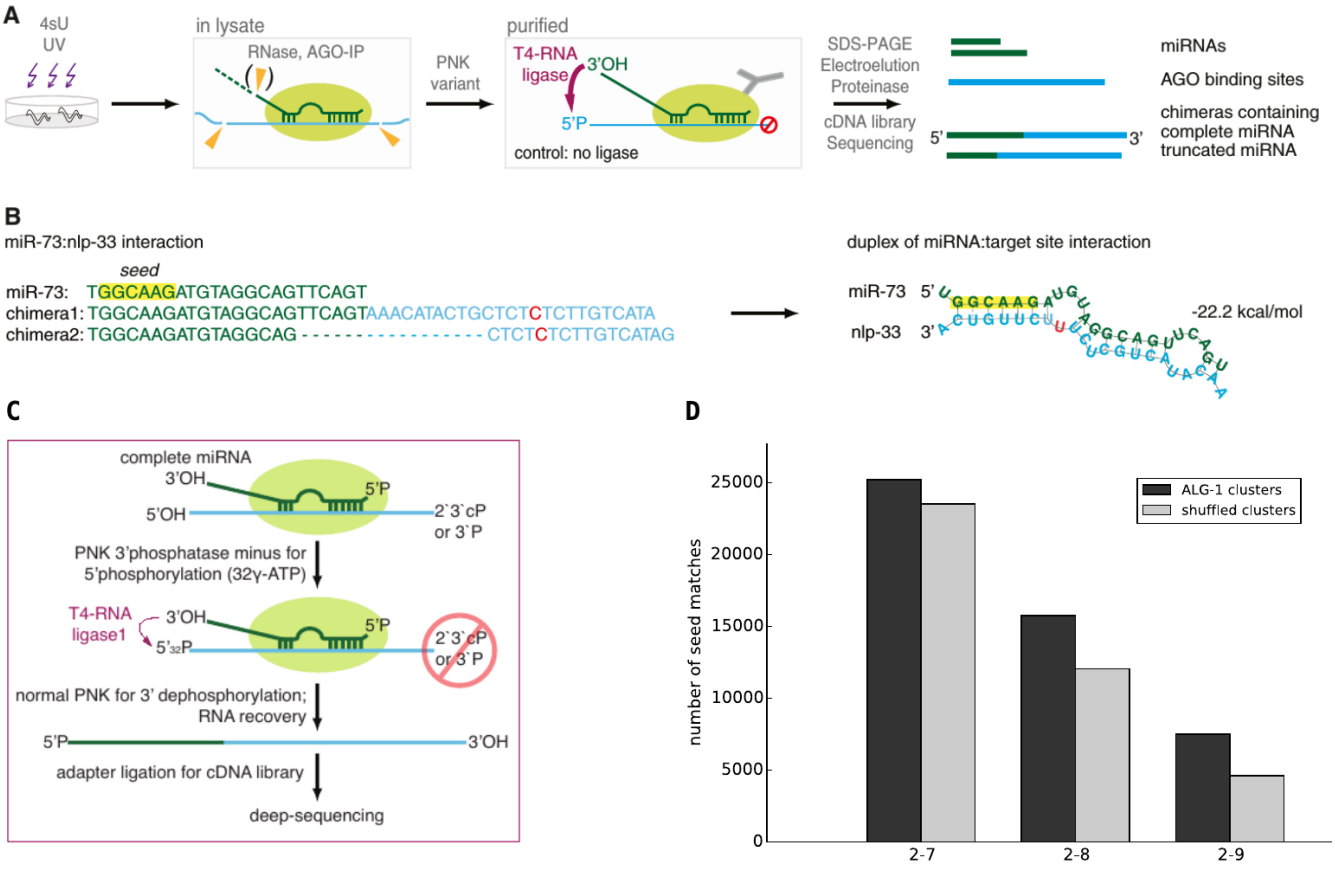

Figure 1| Generation of miRNA:target chimeras via the experimentally added T4 RNA ligase.
(A) *C. elegans* RNA labeled with photoreactive nucleoside 4-thiouridines (4SU) is crosslinked to bound proteins in vivo. After homogenization of worms, the lysate is treated with RNase T1. Some miRNAs are shortened, and others remain complete. Following immunoprecipitation (IP) and washing of AGO, crosslinked RNA is phosphorylated by a PNK variant (leaves 3'ends blocked) and treated with T4 RNA ligase, which ligates the 3'-hydroxyl end of complete miRNAs to bound RNA fragments. Crosslinked RNA is recovered and deep sequenced. Computational analysis detects sequence reads of miRNAs and AGO binding sites, along with chimeric reads containing miRNAs connected to their targets. (Grosswendt et al., 2014)
(B) Example of a miRNA interaction recovered from chimeric reads. Predicted reconstruction of the miRNA:target duplex. Green, miRNA sequence; blue, target sequence; red, T to C conversion. (Grosswendt et al., 2014)
(C) Preparation of RNA ends for generation of chimeras. RNAs crosslinked to washed complexes were radioactively phosphorylated with a PNK enzyme variant that lacks 3'phosphatase activity. Consequently 3'ends of RNAs keep a 2'3'cyclic phosphate (can convert into 3'phosphate). Both 3'end modifications (2'3'cP and 3'P) cannot be ligated by T4 RNA ligase, preventing circularization of target RNAs. As a result, phosphorylated 5'ends of target RNAs are ligated only to the 3'OH ends of complete miRNAs. 3'ends of target RNAs are subsequently converted into hydroxyl groups by the 3'phosphatase activity of PNK, which prepares RNAs for conversion into a cDNA library. (Grosswendt et al., 2014)
(D) Number of seed matches found in ALG-1 binding sites for all C. elegans miRNAs. The cluster was counted as having a seed, if there was at least one match to any miRNA seed.

We decided first to resolve ALG-1 targetome in C. elegans. Since all three protocols are supposed to generate reads coming from miRNA binding sites, we pulled all the libraries together. The sequencing reads were processed in order to remove adapters and nucleotides sequenced with low quality. It resulted in 13.5 million reads with average length of 32nt. These reads were passed to Pipeline for PAR-CLIP Data Analysis (Jens, 2016), a method to recover protein binding sites based on CLIP data. We confidently (FDR<0.05) discovered 29.000 ALG-1 binding sites residing on 8,339 genes. We performed two quality controls on our discoveries. First, we found that the reads covering binding sites had a high T:C conversion rate (14:1 compared to all other possible nucleotide conversions), as expected for PAR-CLIP experiment. Second, the sequences of target regions were enriched with miRNA seed-matches, which are miRNA recognition motifs (Fig. 1D).

Previously 4,806 unique binding sites in 3,093 genes were discovered by Zisoulis and colleagues (Zisoulis et al., 2010). Almost a half of these sites (2.286) are also present in our findings, even though we used worms at different developmental stages. Thus we significantly expanded previously characterized miRNA targetome in C. elegans. Moreover, we managed to lessen the average length of target regions almost thrice compare to Zisoulis et al. (42 vs 122).

3.2 Ligation and control samples contain miRNA:target chimeric reads

We discovered that ~0.18% of the reads coming from ligation samples (modified iPAR-CLIP protocol) were miRNA:target chimeras. As expected, miRNAs resided on the 5'end of chimeric reads, while target part was on the 3'end. There were only few examples of inverted chimeras (that is with target part at 5'end). Most probably they were false discoveries, since our computational pipeline limits false discovery rate to 5%, not to 0%. Since different chimeras can be composed of the same miRNA and overlapping target regions, they were grouped accordingly into objects termed 'interactions'

Surprisingly, we found a comparable fraction of miRNA:target chimeras in the control samples. As the interactions coming from ligation and control samples largely overlapped (Fig. 2A), we set out for further investigation. It turned out that the only difference between chimeras coming from ligation and control samples lay in their composition. In the chimeric reads from the control samples miRNA part was shortened compare to the mature version while the chimeras from the ligation samples were a mixture of cut and complete miRNAs (Fig. 2C). According to the modified iPAR-CLIP protocol only the chimeras with complete miRNAs could be generated. However, the majority of interactions composed of chimeras with complete miRNA are also supported by 'truncated' ones (Fig. 2B). Therefore we did not consider chimeric reads as experimental or computational artifacts and had a closer look at them. We found that guanine is strongly enriched upstream the cleavage site for both

truncated miRNA and target part (Fig. 2D). This observation aligns with the specificity of Rnase T1 used in the protocols, as it prefers to cut upstream guanine. It was reported that 2'3'cyclic phosphate and 5'hydroxyl ends generated by Rnase T1 can be ligated by an endogenous ligation activity (Fig. 2E). Therefore we suggested that chimeras with truncated miRNAs were produced by not yet characterized for C. elegans endogenous ligase and represented almost the same interactions as those produced by experimentally added enzyme.

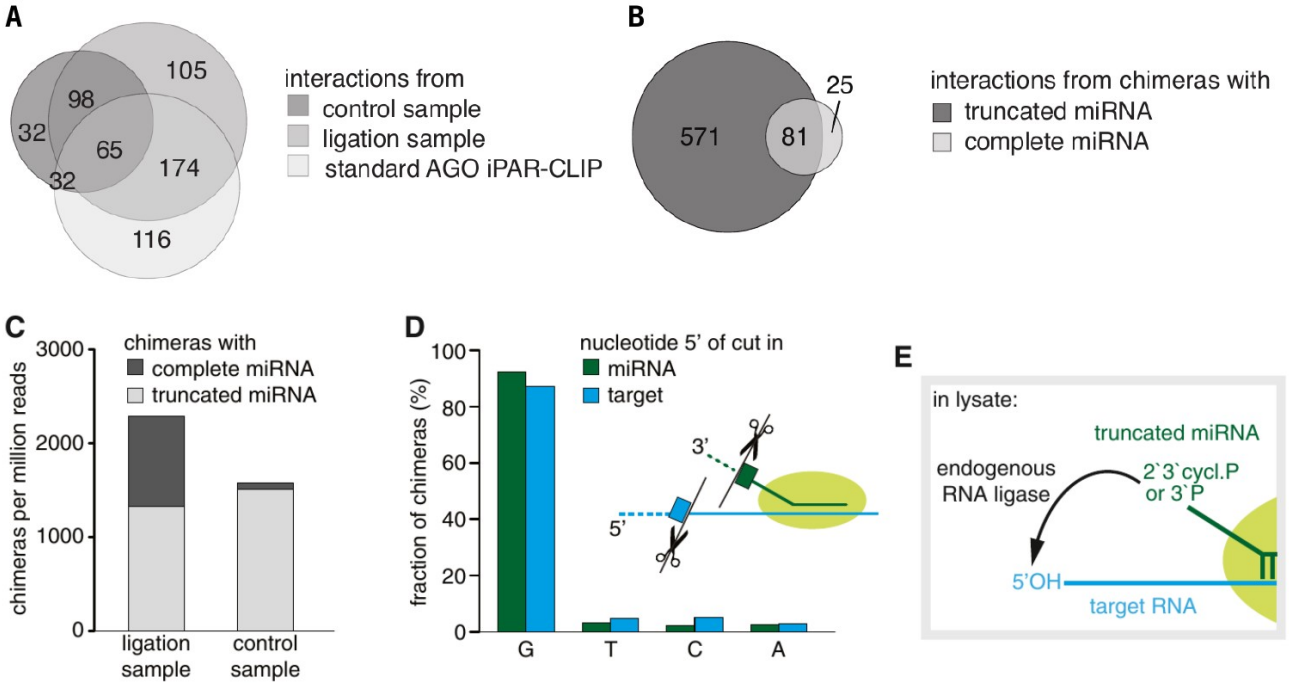

Figure 2| miRNA:target chimeras can be generated via endogenous ligation activity
(A) Chimeras from Modified iPAR-CLIP ligation and control samples and from standard iPAR-CLIP samples recover identical miRNA interactions (only interactions derived from at least 2 chimeric reads were considered). (Grosswendt et al., 2014)
(B) Chimeras with truncated miRNAs (from all C. elegans datasets) and chimeras with complete miRNAs from iPAR-CLIP ligation samples recover identical interactions (only interactions derived from at least 2 chimeric reads were considered). (Grosswendt et al., 2014)
(C) Data from the ligation sample contain chimeras with 30 truncated (length of miRNA sequence R 13 nt) and with complete miRNAs. A comparable fraction of chimeras with truncated miRNAs was also found in a control sample, to which no ligase was added to generate chimeras. (Grosswendt et al., 2014)
(D) miRNA and target ends involved in the ligations of the control sample are highly enriched in an upstream G, suggesting that RNase T1 generated the ends used for this type of ligation. (Grosswendt et al., 2014)
(E) Truncated miRNAs are ligated by the ligase activity of the lysate during IP. (Grosswendt et al., 2014)

Since we found a significant overlap of the interactions coming from different samples and a strong intersection between 'complete' and 'truncated' chimeras, we merged all samples together and analyzed resulting interactions regardless of their origin. In total we discovered 3,348 miRNA:target interactions for C. elegans via mapping to the C. elegans genome.

3.3 Discovered chimeras represent endogenous miRNA:target interactions

Around 90% of the targets in the discovered interactions resided in previously detected ALG-1 binding sites, which suggested convergence between our approach and well-established PAR-CLIP method. Another consequence is that we could map reads to miRNA targetome instead of the whole genome. This strategy helped to narrow down the search space and consequently increase both specificity and sensitivity. Another option was to map to a set of C. elegans 3'UTRs. However this approach did not yield many additional interactions (Fig. 3A). More important, we did not want to narrow miRNA targetome to 3'UTRs only. Indeed, around 70% of the discovered interactions involve 3'UTRs (Fig. 3B). On one hand this finding agrees with known 3'UTR's role as a hub for miRNA binding. On the other hand 30% of interactions would be lost while mapping to the 3' untranslated regions. Interestingly, only 38% ALG-1 binding sites were mapped to 3'UTR's, significantly less than 70% 3'UTR targets in the interactions (Fig. 3C). This difference can be explained by the composition of our samples. Since chimeric reads constitute a small fraction of the total sequencing pool, they may be enriched with strong interactions. Indeed long-lasting interactions have more chances to appear as chimeras and survive all the steps of the protocol. Contrary, the sequencing pool almost exclusively consists of miRNA targets, therefore transient interactions may have enough read support to constitute valid ALG-1 binding site. Thus, chimeric interactions tend to represent strong, long-lasting bindings, while convenient PAR-CLIP provides broader map of miRNA targets.

As it was mentioned above, nucleotide conversion frequencies may serve as a quality control for PAR-CLIP experiments. Remarkably, target parts of chimeric reads appeared to have a high T to C conversion rate (84,3%), 20-fold more than any other kind of nucleotide swap (Fig. 3D). Since T:C conversion is a consequence of crosslink between RNA and protein in PAR-CLIP experiment, we concluded that the detected chimeras indeed represented a physical contact between ALG-1 and targeted RNAs. Interestingly, miRNAs in chimeric reads have T:C conversion frequency lower than conventional miRNA reads (Fig. 3E). It can be explained by a loss of non-crosslinked and non-ligated miRNAs during the protocol, while the ligated ones can survive because of the crosslink in their respective targets.

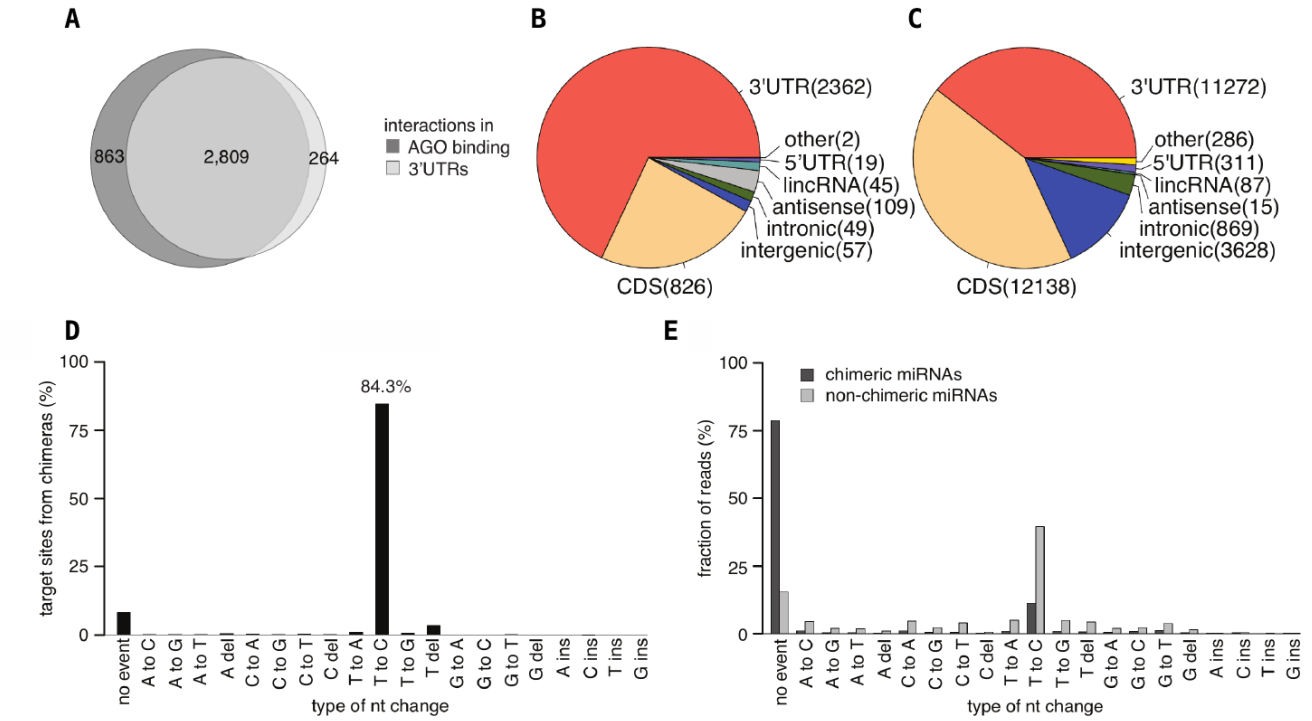

Figure 3| miRNA:target chimeras represent true binding events

(A) The majority (89%) of chimera-derived miRNA target sites (mapped to the transcriptome) overlap with AGO binding sites generated from nonchimeric reads
(B) miRNA target sites derived from chimeras are preferentially located on 3'UTRs
(C) miRNA target sites discovered with a convenient Ago PAR-CLIP are preferentially located on protein coding genes without a skew towards 3'UTRs.
(D) ~84% of sequences ligated to miRNAs have the T to C conversion characteristic for 4sU-crosslinking to bound proteins. deletion (del), insertion (ins)
(E) miRNAs ligated to target RNAs have three-fold less T to C conversions than non-chimeric miRNAs.

One more advantage of using C. elegans is availability of sanity check based on bacterial RNA. Indeed, bacteria are the main food source for the worms. Thus, it was not a surprise that around 30% of sequenced reads had bacterial origin. However, only less than 2% of the interactions were found between C. elegans miRNA and bacterial target, indicating high specificity in detection of chimeras.

Multiple evidences show non-uniform distribution of regulatory elements on miRNA. The seed, nucleotides 2-7 or 2-8 of miRNA, is considered as a main recognition determinant. Binding to 3'end of miRNA also can contribute to the establishing of a relevant interaction. We used RNA co-folding tool RNAhybrid (Rehmsmeier et al., 2004) to calculate a probability for each nucleotide in miRNA to be hybridized with a nucleotide in its target. We discovered a high peak of hybridization frequency for seed region of miRNA, a drop at positions 10-11, and moderate increase for nucleotides in 3'end (Fig. 4A). Two control strategies were used for this analysis: shuffling nucleotides in the target sequences and permuting miRNA:pairs. As expected they both produced Uluru-shaped distributions (uniform distribution with a drop at the edges). Thus, without any prior assumption, we recapitulated well-known rules of miRNA binding based solely on the discovered interactions.

Encouraged by finding an extensive seed usage in the interactions, we decided to have a closer look on recognition via this region. The presence of the following variations of seed binding was checked in the targets: sequence complementarity to the miRNA nucleotides 2-7 (2-7 seed); sequence complementarity to miRNA nucleotides 2-7 with one mismatch (1mm in 2-7) or insertion (1nt bulge in target 2-7); sequence complementarity to the miRNA nucleotides 2-8 with two mismatched nucleotides (2mm in 2-8). More than 80% of the interactions appeared to use a variation of a seed match as recognition motif, far exceeding frequency expected for random sequences or randomly paired miRNA:target couples (Fig. 4B). Hence, the majority of targets were linked to miRNAs in agreement with previously defined binding modes, but not randomly. The rest 20% of 'seedless' interactions showed comparable T:C conversion rate as a total pool, which indicates their association with RISC. One could assume that absence of a seed could be compensated via binding to another part of a miRNA, which will result in the increased basepairing compare to randomly selected sequences. The average free hybridization energy for all miRNA:target pairs was 3.3 kcal/mol lower than for the controls (Fig. 4C), which corresponds to 2-3 paired nucleotides (Mathews et al., 1999; Rehmsmeier et al., 2004). However, for the 'seedless' interactions the difference was very subtle (1.0 kcal/mol, Fig. 4D). Moreover, their hybridization pattern was almost the same as for randomly coupled miRNA:target pairs (Fig. 4E). Thus, we could not find recognition specificity for the seedless interactions. Most probably, they arise from a large amount of weak and transient interactions between RISC and RNAs in the cytoplasm, which were captured with UV-crosslinking and survived all the following experimental steps.

Presence of a seed match in its target is a quality control for miRNA binding. Quality control for PAR-CLIP experiment is a high frequency of T:C conversion. Remarkably, these two sanity checks can be merged for modified iPAR-CLIP. Since crosslink is formed between a nucleotide and an amino acid, it is not plausible to find a T:C conversion inside a seed match hybridized to miRNA seed. Indeed, the frequency of conversion events was very low for the seed matches in the targets compare to the flanks (Fig. 4F). Furthermore, the enrichment of the crosslinks upstream seed match aligns with spatial arrangement of a mRNA inside Argonaute protein (Hafner et al., 2010). Thus, crosslink position in respect to the seed-match confirms the validity of the recovered miRNA:target chimeras.

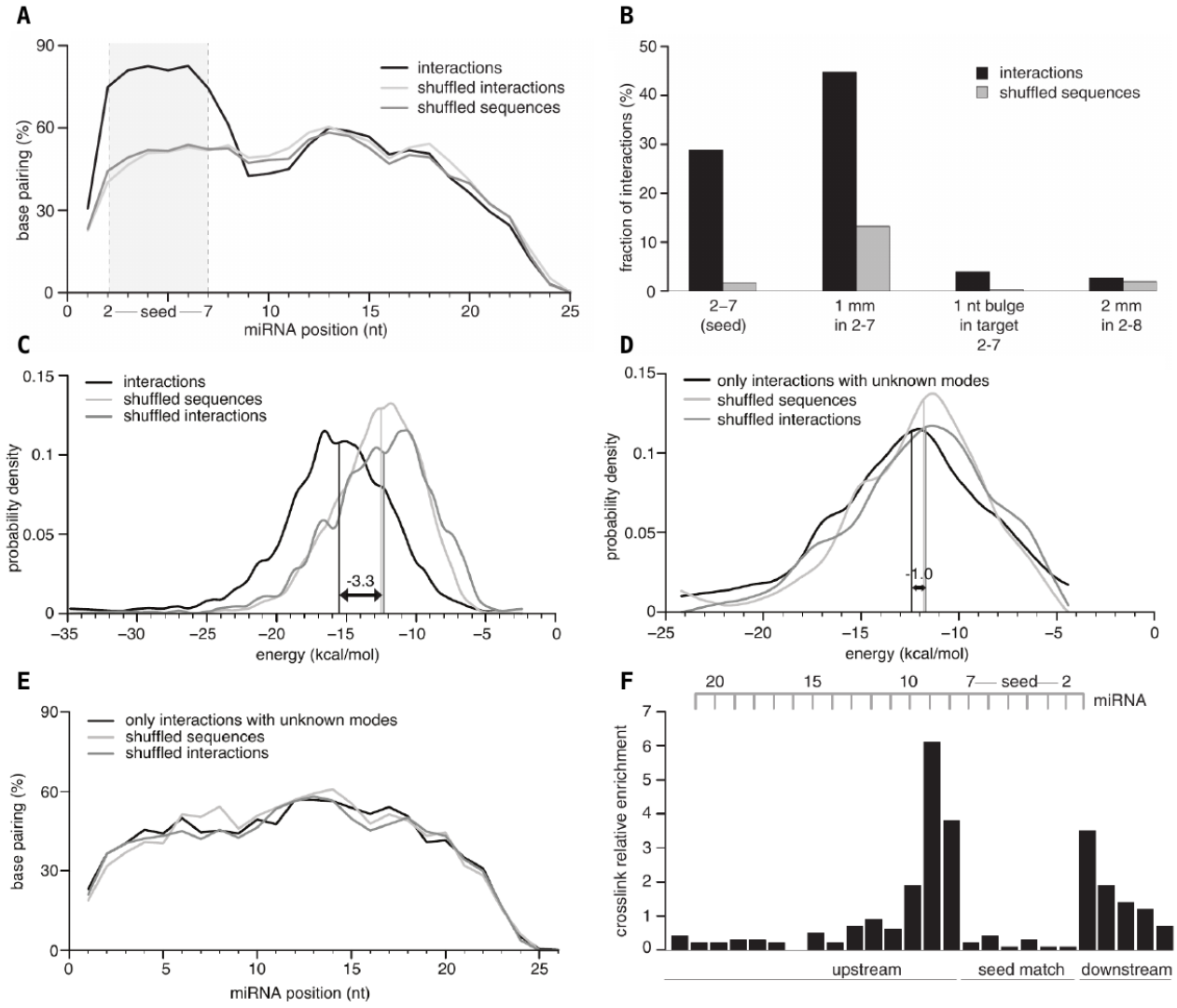

Figure 4| miRNA:target interactions follow expected miRNA targeting- rules

(A) Hybridization profile summarized over all interactions. The predicted frequency of a miRNA position to be base paired is plotted along the miRNA length. Duplex structures of miRNA:targets were predicted by RNAhybrid, allowing G:U pairing. Shuffled sequences (dinucleotides in target sequences are permuted) and shuffled interactions (targets are swapped between miRNAs) served as control.
(B) Target RNAs were analyzed for complementarity to the seed region of their ligated miRNAs. Approximately 80% of interactions possess the tested complementarities. Shuffled sequences (dinucleotides in target sequences are permuted) served as control. mm, mismatch. Mismatches were broadly distributed over all types of nucleotides, including G:U.
(C) Chimera-identified miRNA interactions have a lower binding free energy than expected by chance (Δmedian = 3.3 kcal/mol compared to shuffled interactions). miRNAs and their targets were in silico hybridized by RNAhybrid, allowing G:U pairing.
(D) miRNA:target interactions without any tested binding mode have a lower binding free energy than expected by chance (Δmedian = 1.0 kcal/mol compared to shuffled interactions). miRNAs and their targets were in silico hybridized by RNAhybrid, allowing G:U pairing.
(E) miRNA:target interactions without any tested binding mode show no significant base pairing within the seed. The frequency of a miRNA position to be base paired (as predicted by RNAhybrid) is plotted along the miRNA length. G:U pairings are allowed
(F) Local frequency of crosslink-induced T-to-C conversions in target RNAs from interactions with a perfect 2–7 seed match (normalized to local thymidine frequency). Nucleotides hybridized to the seed of the miRNA are strongly indisposed to crosslink with the protein.

3.4 Thousands of interactions were hidden in published AGO-CLIP datasets

The discovery of miRNA:target chimeras in the PAR-CLIP experiments without ligation step motivated us to explore already published datasets. Indeed, RNase treatment is an unavoidable step in any AGO-CLIP protocol, and endogenous ligation activity is not unique for C. elegans. Therefore, we downloaded raw sequencing reads for all accessible AGO-CLIP experiments and mine them for miRNA:target interactions. It turned out that both PAR-CLIP and HITS-CLIP technologies were able to produce chimeras for human, mouse and virally infected cells. The efficiency of ligation largely differed for the downloaded datasets, which can be explained by variations in experimental setups, biological systems and length of the sequenced reads. The latter was of particular importance, since typical average length of AGO-CLIP reads approximates 30 nt, which is barely enough to confidently identify both miRNA and target part. Therefore the computational pipeline was designed in a way to carefully identify small parts of miRNA (>11nt) and target site (>15nt) (see methods for the details).

Altogether we found 11,000 interactions for human miRNAs, 2,000 for murine miRNAs, 500 for KSHV (Kaposi's sarcoma-associated herpes virus) and 300 for EBV (Epstein-Barr-Virus) miRNAs, respectively (detailed overview of the analyzed datasets can be found in Table 1).

We performed essentially the same analyses for the interactions from downloaded datasets as for the interactions generated by us in C. elegans. Qualitatively we got exactly the same results: a strong enrichment of seed matches for cognate miRNAs in target sequences (Fig. 5A) and typical miRNA binding patterns (Fig. 5B). The observed variations in seed match usage may arise from the differences between species in miRNA targeting and experiment-dependent fraction of the recovered transient interactions.

We further focused on the datasets generated by Kishore and colleagues (Kishore et al., 2011), as they provided the deepest miRNA interaction map for a particular biological system (HEK293 cells). Another advantage was that Kishore and colleagues performed both PAR-CLIP and HITS-CLIP experiments, hence we had a chance to directly compare these approaches. Interactions from PAR-CLIP were of a high quality in terms of T:C conversion rate and position of the crosslink relative to the seed match (Fig. 5 E,F). For HITS-CLIP this check-up was impossible to perform, since there is no particular type of conversion for this method. PAR-CLIP and HITS-CLIP showed similar binding patterns and hybridization energy distributions (Fig. 5C,D). However, interactions coming from HITS-CLIP methodology used pairing via perfect seed more frequently. Noteworthy, PAR-CLIP interactions compensated the gap in perfect seed usage with more frequent pairing via imperfect seeds.

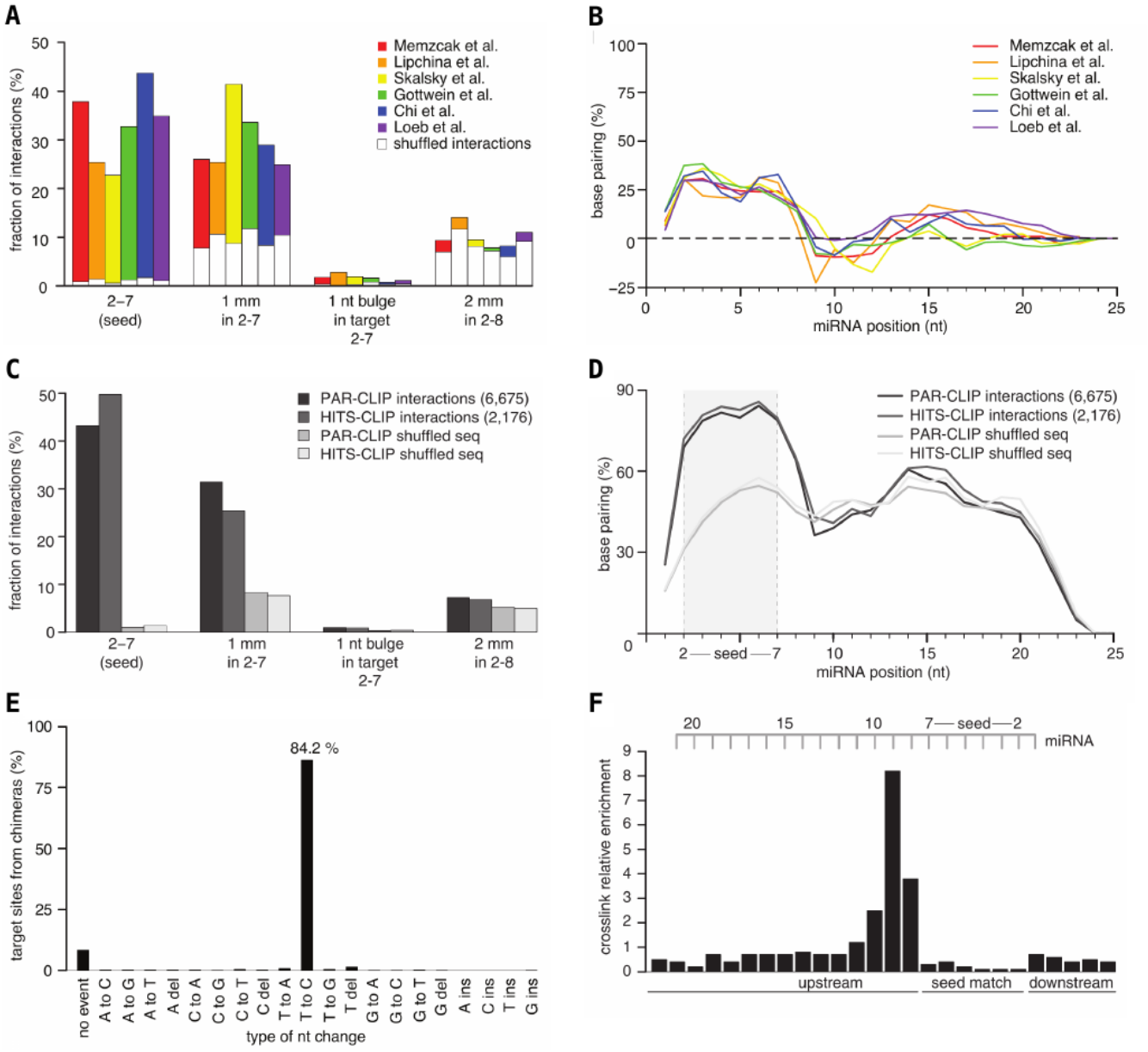

Figure 5| miRNA:target interactions recovered in AGO-CLIP experiments in various species show hallmarks of being genuine.

(A) Bindings modes usage among the discovered miRNA:target interactions in different datasets; shuffled interactions (white) served as control.
(B) Hybridization profiles of interactions from all datasets examined in the present study; controls (shuffled target sequences) were subtracted. RNAhybrid predicted, G:U allowed.
(C) Bindings modes usage among the discovered miRNA:target interactions in AGO2 PAR-CLIP and HITS-CLIP data (HEK293 cells) from Kishore et al. (2011); shuffled interaction served as control.
(D) Hybridization profiles summarized over all interactions found in Kishore AGO2 PAR-CLIP and HITS-CLIP datasets.
(E) ~84% of target sequences ligated to miRNAs in human AGO PAR-CLIP data (Kishore et al.) have a T to C conversion characteristic for 4sU-crosslinking to bound proteins.
(F) Local frequency of crosslink-induced T-to-C conversions in target RNAs from interactions with a perfect 2–7 seed match (normalized to local thymidine frequency). Positions hybridized to the seed of the miRNA are strongly indisposed to crosslink.

Table 1. miRNA:targets from re-analysis of published AGO CLIP data

publication	CLIP	RNase	Argonaute	model system	miRNA:targets	FDR %
Kishore *et al.,* 2011	PAR	T1	Ago2	human cells (HEK 293)	6675	4
	HITS	T1	Ago2	human cells (HEK 293)	2176	3
Memczak *et al.,* 2013	PAR	T1	Ago1	human cells (HEK 293)	1010	4
Lipchina *et al.,* 2011	PAR	T1	Ago2	human embryonic stem cells	146	5
Skalsky *et al.,* 2012	PAR	T1	Ago2	EBV-infected lymphoblastoid cell lines	74 viral 997 human	3
Gottwein *et al.,* 2011	PAR	T1	Ago2	primary effusion lymphoma cell lines (BC-1, BC-3)	660 viral 236 human	4
Chi *et al.,* 2009	HITS	A	Ago	mouse brain	565	4
Loeb *et al.,* 2012	HITS	A	Ago2	mouse T-cells WT	1269	4
	HITS	A	Ago2	mouse T-cells mir-155 KO	260	4

Raw sequencing data from listed AGO PAR-CLIP and HITS-CLIP datasets contain miRNA:target chimeras. BC-1 and BC-3 are primary effusion lymphoma derived cell lines infected with Epstein-Barr-Virus (EBV) and Kaposi's sarcoma-associated herpes virus (KSHV). (Grosswendt et al., 2014)

3.5 miRNA seed matches are selected in course of evolution

Discovered interactions followed known rules of miRNA binding and passed the quality controls for CLIP experiments. But do they represent functionally important regulatory events? As it was discussed in the introduction section, the most straightforward way to check for functional importance of a stretch of nucleotides is to assess its conservation. In the case of miRNA:target interactions this stretch of nucleotides is a seed match to the cognate miRNA on target sequence. Since we consider 'chimeric' approach as an enhancement to AGO-IP methods, we decided to compare conservation of the seed matches inside AGO binding sites to the seed matches defined by miRNA:target interactions. Seed matches found in 3'UTRs served as another control, which in some sense was a comparison of our method with a number of bioinformatics tools, which usually focus on miRNA seed matches in 3'UTRs. Since conservation of binding sites might depend on how well is conserved the corresponding miRNA, we split out analysis for the groups of miRNAs with identical seed.

Conservation of seed-matches defined by the discovered interactions appeared to be significantly higher for almost all miRNA families for both worm and human (Fig. 6A). Moreover an enhanced conservation was observed for the seed matches with one mismatch (Fig. 6B). As far as we know, it was the first large-scale evidence that imperfect seed pairing is relevant in post-transcriptional regulation. We assumed that the main reason why seed matches pooled from miRNA:target interactions appeared to be more conserved, is that for 3'UTRs and AGO binding sites we had to predict binding to miRNA. Since the probability of a seed-match 6mer to appear randomly in a genome is 1 per 4000 nucleotides, the sequence-based predictions might be contaminated with false positives. On contrast, for miRNA:targets we knew exactly which miRNA binds the region of interest and did not have to make a guess. It is one of the examples how 'ligation' approach provides a new resolution of miRNA interaction map.

Binding via imperfect seed explains a considerable fraction of miRNA:target pairs. Since this kind of binding is impossible to specifically predict with pure computational methods and almost impossible for AGO-CLIP based approaches, we had a chance to explore imperfect seed pairing features for the first time. We found that a bulge inside imperfect duplex tends to reside at the edges of seed. We found this tendency for all the species analyzed, which suggest its generality for miRNA targeting (Fig. 6C,D). Thus, nucleotides 3-6 in miRNA 5'end form a 'core' seed, which can be explained by thermodynamical reasons. Indeed, duplex involving core seed can still benefit from staking interactions, while internal bulge would disturb them.

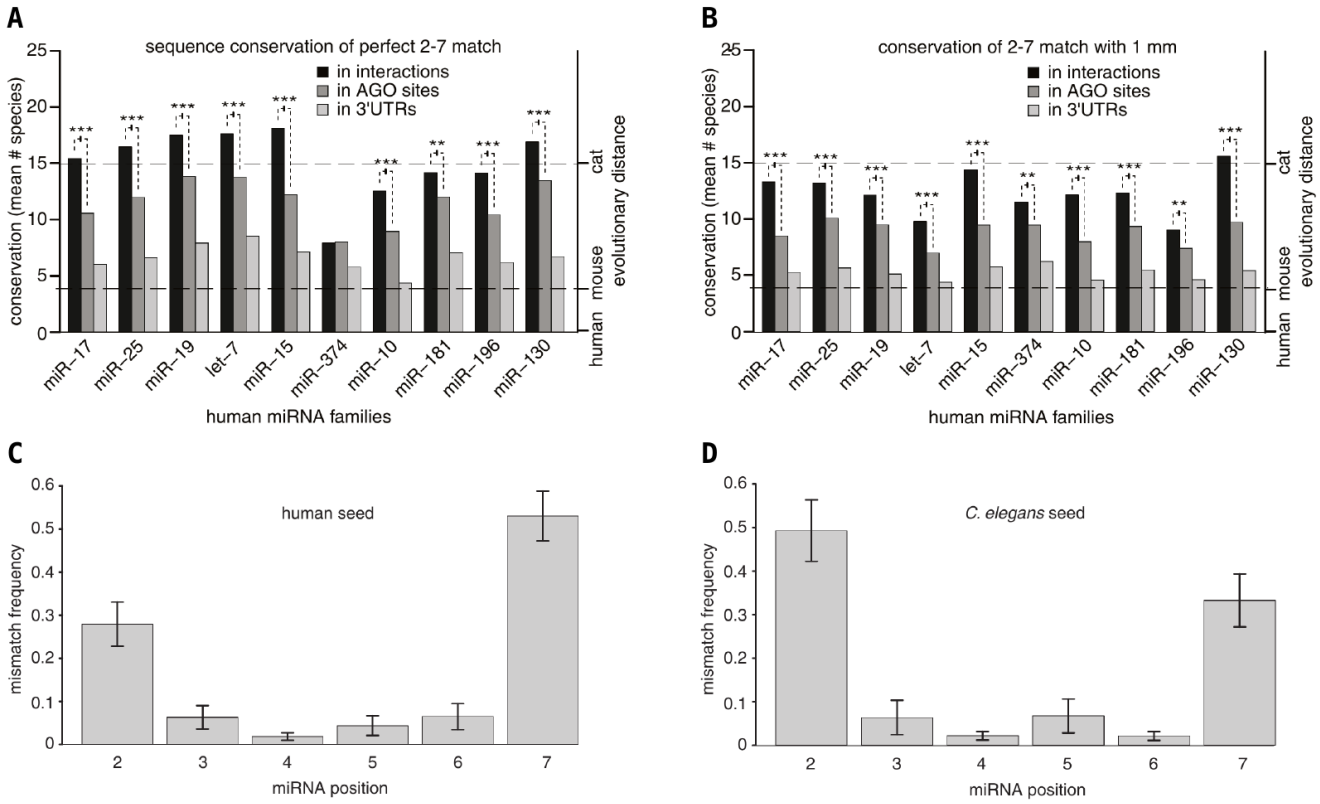

Figure 6| miRNA:target interactions are well-conserved in various species

(A and B) Conservation across 31 vertebrate species of perfect seed (2–7) matches (A) and seed matches with 1 nt mismatch (1 mm) (B) from human miRNA: targets recovered by analysis of chimeras. Conservation of other seed matches for the same miRNA served as a control. A perfect seed match in human was counted as conserved if present at the same position in the alignment. A seed match with 1 mm was deemed conserved if the identical 1 mm seed match or the perfect seed match was present at the same position in the alignment. On average, 100 miRNA interactions (median) were included per miRNA family. miRNA:targets with a mismatch in the 2–7 seed were significantly conserved (***$p < 0.005$; **$p < 0.01$, Mann-Whitney U test), but to a lower degree than perfect seed matches.
(C) Mismatches in seed sites occur predominantly at position 2 or 7 of the miRNA. Shown is the positional mismatch frequency for interactions with a 2–7 match containing 1 mismatch, averaged over different miRNA families.
(D) As in (C), but in C. elegans.

3.6 miRNA:target chimeras allow to distinguish miRNA family members

The information on direct miRNA:target interactions allows to distinguish targets of miRNAs sharing the same seed sequence, while computational predictions, even augmented with AGO-CLIP, cannot distinguish between them. Therefore we had a chance to analyze the interrelations between miRNA family members for the first time. As expected we found that the members of miRNA families tend to have common binding sites(Fig. 7A,B). However, the majority of targets were assigned to a single miRNA from a family. It can be caused by a relative shallowness of chimeric data, but also by contribution of miRNA 3'end binding. Indeed, members of the same family vary in their 3'end sequence which can cause differential targeting. Based on chimeric interactions we were able to explore a role of 3'end binding in miRNA:target pairing. Targets of 18 human miRNAs showed a significant complementarity for the 3'ends (Table 2). Remarkably, members of hsa-miR-196 family have exactly the same sequence but differ in 12th nucleotide, and consequently vary in 3'end binding. MiR-196a typically pairs via nucleotides 13-19, while miR-196b extends complementarity to the 12th nucleotide. Thus, these two almost identical guide RNAs differ in affinity to their targets. The same phenomenon was observed for hsa-miR-15 family. MiRNA-15a and miRNA-15b typically pairs via 12-17 and 13-17 bases respectively, while the 12th and three last nucleotides is only variation between them.

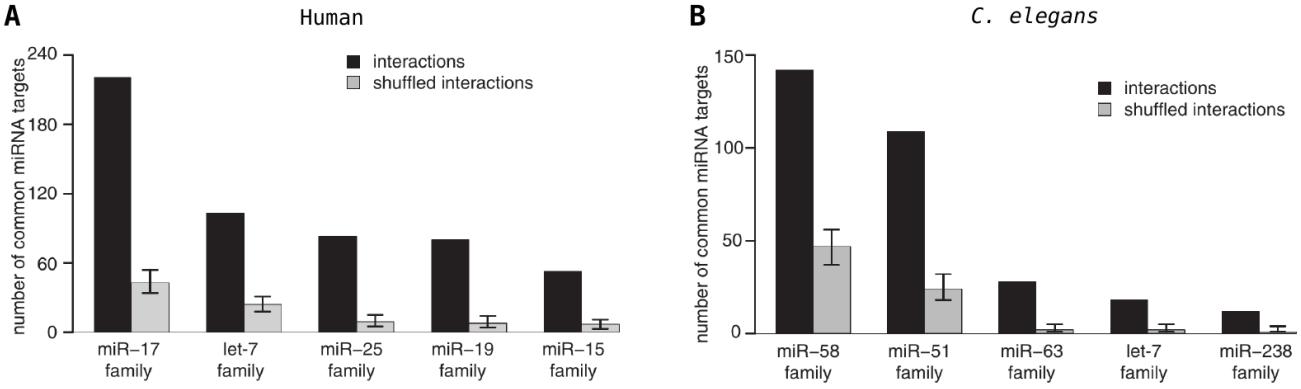

Figure 7| Members of miRNA families tend to share common binding sites

(A and B) Individual target sites in human (A) and *C.elegans* (B) are ligated to members of the same miRNA family more often than expected by chance.

Table 2. Sequence complementarities found in 3'part of miRNAs

miRNA ID	miRNA sequence	nucleotide positions with complementarity (empirical p<0.01)
hsa-miR-10a	TACCCTGTAGATCCGAATTTGTG	19-22
hsa-miR-19a	TGTGCAAATCTATGCAAAACTGA	12-18
hsa-miR-19b	TGTGCAAATCCATGCAAAACTGA	12-17
hsa-miR-30e	TGTAAACATCCTTGACTGGAAG	13-16
hsa-miR-17	CAAAGTGCTTACAGTGCAGGTAG	12-18
hsa-miR-20a	TAAAGTGCTTATAGTGCAGGTAG	13-18
hsa-miR-196a	TAGGTAGTTTCATGTTGTTGGG	13-19
hsa-miR-196b	TAGGTAGTTTCCTGTTGTTGGG	12-19
hsa-miR-92a	TATTGCACTTGTCCCGGCCTGT	11-16
hsa-miR-106b	TAAAGTGCTGACAGTGCAGAT	12-18
hsa-miR-16	TAGCAGCACGTAAATATTGGCG	13-17
hsa-miR-33a	GTGCATTGTAGTTGCATTGCA	12-21
hsa-miR-93	CAAAGTGCTGTTCGTGCAGGTAG	15-18
hsa-miR-15a	TAGCAGCACATAATGGTTTGTG	12-17
hsa-miR-15b	TAGCAGCACATCATGGTTTACA	13-17
hsa-miR-221	AGCTACATTGTCTGCTGGGTTTC	12-17
hsa-miR-423-3p	AGCTCGGTCTGAGGCCCCTCAGT	12-22
hsa-miR-3168	GAGTTCTACAGTCAGAC	10-15

Human miRNAs recovered from chimera analysis of CLIP data by Kishore et al., 2011 with significant complementarities between their 3'part and ligated targets. Dinucleotide shuffling of target sequences served as control.

3.7 Discovered miRNA:target interactions arise from relevant regulatory events

Experimental discovery of a targetome for a particular miRNA typically employs the following strategy. MiRNA expression is perturbed; it can be a knockout (KO), knockdown (KD) or overexpression. Then copy numbers of mRNA transcripts are quantified and contrasted to their wild-type expression. The genes which are significantly affected by miRNA perturbation are assumed to be its targets. The miRTarBase (Chou et al., 2016) comprises a collection of miRNA:target pairs experimentally confirmed around the world. Therefore we decided to use this information to confirm the functional relevance of our findings. 148 unique miRNA:gene pairs discovered by us were also found in miRTarBase, far exceeding an overlap expected at random ($p < 0.0001$, Table 3). These include famous, well-studied interactions such as let-7:DICER, let7:lin41, lin4:daf-12, lin4:lin-28, miR-196:HOX genes..

Further, we set out to directly compare our findings with several published miRNA perturbation experiments. Hafner and colleagues (Hafner et al., 2010) inhibited the top 25 expressed miRNAs in HEK293 cell line and track subsequent changes in global gene expression. For target genes assigned

by our analysis to these 25 miRNAs we observed a significant shift in the transcripts copy numbers (Fig. 8B, $p < 7.63 \times 10^{-25}$). Remarkably, 590 out of 1,115 interactions involving the perturbed miRNAs do not use pairing via perfect seed match, hence could not be predicted by computational means. Essentially the same results were obtained for the dataset produced by Lipchina and colleagues (Lipchina et al., 2011). The authors inhibited miR-367 and the members of miR-302 family in human embryonic stem cells. The targets for these miRNAs turned out to be more derepressed than the whole pool of genes (Fig. 8C, $p < 9.8 \times 10^{-8}$; KS test), even though the majority (41 out of 68) of them lacked a seed match. Differential AGO-CLIP was also performed for a mouse cell line. Loeb and colleagues (Loeb et al., 2012) measured miRNA targetome in embryonic stem cells lacking MIR-155 genes versus wild-type system. We were able to discover 46 interactions involving miR-155 in the WT sample and no interactions in KO. Consistently, the read coverage on the miR-155 binding sites for wild-type appeared to be significantly higher than for the knockout (Fig. 8A). Even the targets with no seed match (25 out of 46) were globally down-regulated ($P<0.004$, Kolmogorov-Smirnov [KS] test), proving their functionality in post-transcriptional gene control. Finally, we assessed the influence of miRNA targeting on protein level. We looked at protein levels changes in pulsed SILAC data after miR-155 overexpression in HeLa cells (Selbach et al., 2008). Chimera analysis revealed 91 target genes for miR-155 which were also measured by Selbach and colleagues. Protein synthesis of these targets appeared to be significantly downregulated (Fig. 8D). Thus, the functional relevance of our findings was also observed on protein level.

In summary, the biological importance of miRNA:target interactions found via chimeric approach was confirmed by multiple perturbation experiments. For all the data analyzed, we observed a significant downregulation for the targets associated with the overexpressed miRNAs and upregulation for those associated with the inhibited ones.

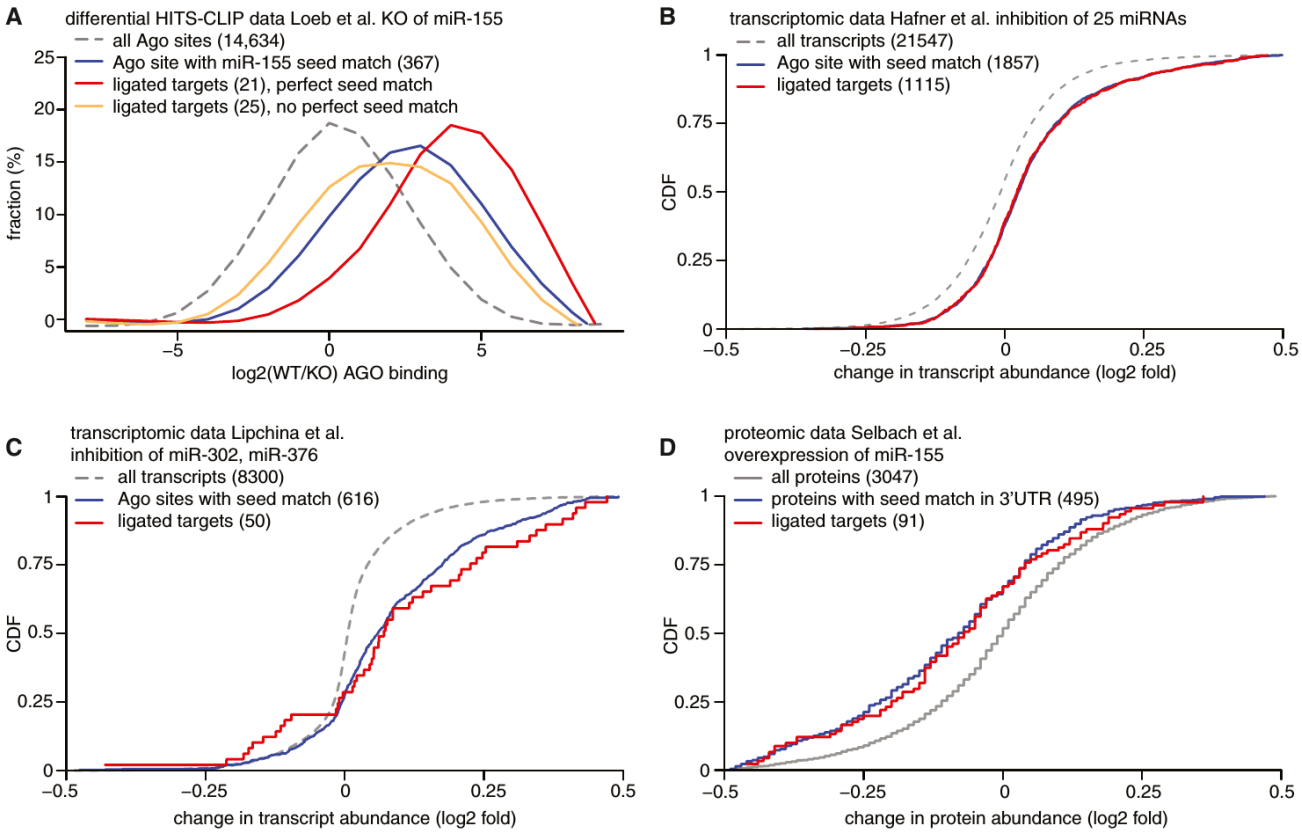

Figure 8| Functional miRNA targets derived from chimeras

(A) HITS-CLIP sequencing data from WT and miR-155 KO cells (Loeb et al., 2012) were analyzed for chimeras containing miR-155. miR-155 ligated target sites with a perfect 2–7 seed match (red) were targeted by AGO2 in WT cells more often than in miR-155 KO cells compared to all transcripts (dashed) and all clusters with a seed match (blue) ($p < 0.004$; KS test). miR-155 ligated target sites without a perfect 2–7 match (orange) are AGO2-bound in WT significantly more often than in KO cells ($p < 0.003$; KS test).

(B–D) miRNA perturbation data demonstrate functionality of chimera-identified miRNA interactions. Changes in transcript abundance after inhibition of 25 miRNAs in HEK293 cells (Hafner et al., 2010) (B) and miR-302a/b/c/d, miR-367 in mouse embryonic stem cells (Lipchina et al., 2011) (C), and changes in protein abundance after overexpression of miR-155 in a human cell line (Selbach et al., 2008) (D). Targets recovered in chimeras with these miRNAs (from all HEK293 data and human embryonic stem cell data, respectively; Table 1) were upregulated upon miRNA inhibition on the transcript level (B and C) and downregulated on the protein level upon miRNA overexpression (D).

Table 3. Recovery of previously verified miRNA:target interactions

miRNA ID	Gene symbol	Dataset
hsa-let-7a	DICER1	Kishore et al., 2011 HITS-CLIP and PAR-CLIP
hsa-let-7a	FOXA1	Kishore et al., 2011 PAR-CLIP
hsa-let-7a	HMGA2	Kishore et al., 2011 PAR-CLIP
hsa-let-7a	IGF2BP1	Kishore et al., 2011 PAR-CLIP
hsa-let-7a	ZFP36L1	Kishore et al., 2011 PAR-CLIP
hsa-let-7b	AKAP8	Kishore et al., 2011 PAR-CLIP
hsa-let-7b	AURKB	Kishore et al., 2011 PAR-CLIP
hsa-let-7b	DICER1	Kishore et al., 2011 HITS-CLIP and PAR-CLIP
hsa-let-7b	HMGA2	Kishore et al., 2011 PAR-CLIP
hsa-let-7b	IGF2BP1	Kishore et al., 2011 PAR-CLIP
hsa-let-7b	PDE12	Kishore et al., 2011 PAR-CLIP
hsa-let-7b	RDH10	Kishore et al., 2011 PAR-CLIP
hsa-let-7b	SPRYD4	Kishore et al., 2011 PAR-CLIP
hsa-let-7c	DICER1	Kishore et al., 2011 HITS-CLIP and PAR-CLIP
hsa-let-7c	HMGA2	Kishore et al., 2011 PAR-CLIP
hsa-let-7d	HMGA2	Kishore et al., 2011 PAR-CLIP
hsa-let-7g	IGF2BP1	Kishore et al., 2011 HITS-CLIP and PAR-CLIP
hsa-miR-101	ARID1A	Kishore et al., 2011 PAR-CLIP
hsa-miR-101	ATP5B	Kishore et al., 2011 PAR-CLIP
hsa-miR-106b	CCND1	Kishore et al., 2011 PAR-CLIP
hsa-miR-106b	E2F1	Kishore et al., 2011 HITS-CLIP and PAR-CLIP
hsa-miR-106b	ITCH	Kishore et al., 2011 PAR-CLIP
hsa-miR-10b	BCL2L11	Kishore et al., 2011 PAR-CLIP
hsa-miR-10b	HOXD10	Kishore et al., 2011 PAR-CLIP
hsa-miR-124	SLC16A1	Kishore et al., 2011 PAR-CLIP
hsa-miR-151	ARHGDIA	Kishore et al., 2011 PAR-CLIP
hsa-miR-155	RCN2	Kishore et al., 2011 PAR-CLIP
hsa-miR-155	VAMP3	Kishore et al., 2011 PAR-CLIP
hsa-miR-15a	VEGFA	Kishore et al., 2011 HITS-CLIP and PAR-CLIP
hsa-miR-15b	CCNE1	Kishore et al., 2011 PAR-CLIP
hsa-miR-15b	VEGFA	Kishore et al., 2011 PAR-CLIP
hsa-miR-16	ALG3	Kishore et al., 2011 PAR-CLIP
hsa-miR-16	CCNT2	Kishore et al., 2011 PAR-CLIP
hsa-miR-16	SHOC2	Kishore et al., 2011 PAR-CLIP
hsa-miR-17	APP	Kishore et al., 2011 PAR-CLIP
hsa-miR-17	CCND1	Kishore et al., 2011 HITS-CLIP and PAR-CLIP
hsa-miR-17	E2F1	Kishore et al., 2011 HITS-CLIP
hsa-miR-17	E2F3	Kishore et al., 2011 HITS-CLIP
hsa-miR-17	NPAT	Kishore et al., 2011 HITS-CLIP
hsa-miR-17	PKD2	Kishore et al., 2011 HITS-CLIP
hsa-miR-17	TGFBR2	Kishore et al., 2011 PAR-CLIP
hsa-miR-17	WEE1	Memczak et al., 2013 PAR-CLIP
hsa-miR-17	ZNFX1	Kishore et al., 2011 HITS-CLIP and PAR-CLIP
hsa-miR-181a	GATA6	Kishore et al., 2011 HITS-CLIP
hsa-miR-181a	PLAG1	Kishore et al., 2011 PAR-CLIP
hsa-miR-193b	CCND1	Kishore et al., 2011 HITS-CLIP

miRNA ID	Gene symbol	Dataset
hsa-miR-196a	CDKN1B	Kishore et al., 2011 PAR-CLIP
hsa-miR-196a	HOXB7	Kishore et al., 2011 PAR-CLIP
hsa-miR-196a	HOXB8	Kishore et al., 2011 HITS-CLIP
hsa-miR-196a	HOXC8	Kishore et al., 2011 HITS-CLIP
hsa-miR-196a	HOXD8	Kishore et al., 2011 HITS-CLIP and PAR-CLIP
hsa-miR-196b	HOXB8	Kishore et al., 2011 HITS-CLIP
hsa-miR-196b	HOXC8	Kishore et al., 2011 PAR-CLIP
hsa-miR-19a	SMAD4	Kishore et al., 2011 PAR-CLIP
hsa-miR-19b	ARID4B	Kishore et al., 2011 PAR-CLIP
hsa-miR-19b	MYLIP	Kishore et al., 2011 PAR-CLIP
hsa-miR-20a	BCL2	Kishore et al., 2011 PAR-CLIP
hsa-miR-20a	CCND1	Kishore et al., 2011 HITS-CLIP and PAR-CLIP
hsa-miR-20a	E2F1	Kishore et al., 2011 PAR-CLIP
hsa-miR-20a	EGLN3	Kishore et al., 2011 HITS-CLIP and PAR-CLIP
hsa-miR-221	CDKN1B	Memczak et al., 2013 PAR-CLIP, Kishore et al., 2011 PAR-CLIP
hsa-miR-221	DDIT4	Memczak et al., 2013 PAR-CLIP
hsa-miR-222	PPP2R2A	Kishore et al., 2011 PAR-CLIP
hsa-miR-24	CDKN1B	Memczak et al., 2013 PAR-CLIP, Kishore et al., 2011 HITS-CLIP
hsa-miR-25	BCL2L11	Kishore et al., 2011 HITS-CLIP and PAR-CLIP
hsa-miR-26a	CDC6	Kishore et al., 2011 PAR-CLIP
hsa-miR-26a	GSK3B	Memczak et al., 2013 PAR-CLIP
hsa-miR-27a	SPRY2	Kishore et al., 2011 PAR-CLIP
hsa-miR-27a	THRB	Kishore et al., 2011 PAR-CLIP
hsa-miR-29a	MCL1	Kishore et al., 2011 PAR-CLIP
hsa-miR-29a	PIK3R1	Kishore et al., 2011 PAR-CLIP
hsa-miR-29b	HMGA2	Kishore et al., 2011 PAR-CLIP
hsa-miR-29b	MCL1	Kishore et al., 2011 PAR-CLIP
hsa-miR-29c	MCL1	Kishore et al., 2011 PAR-CLIP
hsa-miR-424	WEE1	Kishore et al., 2011 PAR-CLIP
hsa-miR-7	CNN3	Memczak et al., 2013 PAR-CLIP
hsa-miR-7	CNOT8	Kishore et al., 2011 PAR-CLIP
hsa-miR-7	PSME3	Kishore et al., 2011 PAR-CLIP
hsa-miR-92a	BCL2L11	Memczak et al., 2013 PAR-, Kishore et al., 2011 HITS- and PAR-CLIP
hsa-miR-93	E2F1	Kishore et al., 2011 PAR-CLIP
hsa-miR-93	TP53INP1	Memczak et al., 2013 PAR-CLIP
hsa-miR-7	CDR1AS*	Kishore et al., 2011 HITS-CLIP and PAR-CLIP
hsa-miR-155	MSI2	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	EHD1	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	IKBIP	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	SMAD5	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	MYO1E	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	SMAD2	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	RCOR1	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	KBTBD2	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	BACH1	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	NARS	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	CBFB	Skalsky et al., 2012 PAR-CLIP

miRNA ID	Gene symbol	Dataset
hsa-miR-155	VEZF1	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	PHC2	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	TXNRD1	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	ARID2	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	EDEM3	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	TRIP13	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	PAPOLA	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	VAMP3	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	TAB2	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	ERMP1	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	TRAM1	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	POLE3	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	GNA13	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	CKAP5	Skalsky et al., 2012 PAR-CLIP
hsa-miR-34a	VAMP2	Skalsky et al., 2012 PAR-CLIP
hsa-miR-142	RAC1	Skalsky et al., 2012 PAR-CLIP
hsa-miR-181a	BCL2	Skalsky et al., 2012 PAR-CLIP
hsa-miR-21	BCL2	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	DAZAP2*	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	FOS*	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	CLIC4*	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	RFK*	Skalsky et al., 2012 PAR-CLIP
hsa-miR-155	CDKN1A*	Skalsky et al., 2012 PAR-CLIP
hsa-let-7f	PRDM1	Gottwein et al., 2011 PAR-CLIP
hsa-miR-101	ARID1A	Gottwein et al., 2011 PAR-CLIP
hsa-miR-29a	MCL1	Gottwein et al., 2011 PAR-CLIP
hsa-miR-29b	MCL1	Gottwein et al., 2011 PAR-CLIP
ebv-miR-BART1	LY75*	Skalsky et al., 2012 PAR-CLIP
ebv-miR-BHRF1	LY75*	Skalsky et al., 2012 PAR-CLIP
kshv-miR-K12-1	RAD21*	Gottwein et al., 2011 PAR-CLIP
kshv-miR-K12-4	YWHAB*	Gottwein et al., 2011 PAR-CLIP
kshv-miR-K12-11	BACH1	Gottwein et al., 2011 PAR-CLIP
mmu-miR-106a	Stat3	Loeb et al., 2012 HITS-CLIP
mmu-miR-124	Arfip1	Chi et al., 2009 HITS-CLIP
mmu-miR-124	Cd164	Chi et al., 2009 HITS-CLIP
mmu-miR-124	Ptbp1	Chi et al., 2009 HITS-CLIP
mmu-miR-150	Myb	Loeb et al., 2012 HITS-CLIP
mmu-miR-155	Jarid2	Loeb et al., 2012 HITS-CLIP
mmu-miR-155	Lpin1	Loeb et al., 2012 HITS-CLIP
mmu-miR-155	Trib1*	Loeb et al., 2012 HITS-CLIP
mmu-miR-155	Zc3h11a*	Loeb et al., 2012 HITS-CLIP
mmu-miR-17	Rbl2	Loeb et al., 2012 HITS-CLIP
mmu-miR-17	Stat3	Loeb et al., 2012 HITS-CLIP
mmu-miR-23a	Lmnb1	Loeb et al., 2012 HITS-CLIP
mmu-miR-23b	Lmnb1	Loeb et al., 2012 HITS-CLIP
mmu-miR-24	Bcl2l11	Loeb et al., 2012 HITS-CLIP
mmu-miR-27a	Runx1	Loeb et al., 2012 HITS-CLIP

miRNA ID	Gene symbol	Dataset
mmu-miR-29a	Dnmt3a	Loeb et al., 2012 HITS-CLIP
mmu-miR-33	Abca1	Chi et al., 2009 HITS-CLIP
cel-let-7	daf-12	Grosswendt et al., 2014 iPAR-CLIP
cel-let-7	hbl-1	Grosswendt et al., 2014 iPAR-CLIP
cel-let-7	T14B1.1*	Grosswendt et al., 2014 iPAR-CLIP
cel-let-7	Lin-28	Grosswendt et al., 2014 iPAR-CLIP
cel-let-7	Lin-41	Grosswendt et al., 2014 iPAR-CLIP
lin-4	Lin-28	Grosswendt et al., 2014 iPAR-CLIP

149 of miRNA targets identified by analysis of chimeric reads were previously reported and verified by others; 136 of them are annotated in miRTarbase; miRNA targets not listed in miRTarBase are marked with (*) (CDR1as Memczak et al., 2013; Trib1 and Zc3h11a Loeb et al., 2012; T14B1.1 Lall et al., 2006; DAZAP2, CLIC4, CDKN1A, LY75 Skalsky et al., 2007; FOS, RFK, RAD21, YWHAB Gottwein et al., 2007)

3.8 Analysis of ligation products unambiguously revealed targets of viral miRNAs

The direct identification of miRNA:targets is specifically important for the study of viral miRNA. Indeed, some human pathogens, including Kaposi sarcoma herpesvirus (KSHV) and Epstein-Barr virus, encode for a number of miRNAs (Grundhoff et al., 2006; Kincaid and Sullivan, 2012; Pfeffer et al., 2004; Samols et al., 2005). These viruses infect only humans, hence binding sites for their miRNAs are not necessarily evolutionary conserved, which complicates their identification. Moreover some of viral miRNAs share seed sequence with human ones, making impossible an unambiguous assignment of a particular target to a miRNA identity (Gottwein et al., 2007, 2011; Manzano et al., 2013; Skalsky et al., 2007). One of the examples is KSHV miR-K11, which has exactly the same first 8 nucleotides as human miR-155 (Fig. 9A), and recapitulates miR-155 oncogenic properties in B cells (Boss et al., 2011; Dahlke et al., 2012; Linnstaedt et al., 2010; Skalsky et al., 2007). Consistently, we found 11 binding sites shared between miR-155 and miR-K11 (Fig. 9A) in AGO2-CLIP data from KSHV-infected B cells (Gottwein et al., 2011) and lymphoblastoid cells (Skalsky et al., 2012)The number of common targets was significantly greater than expected to appear by chance (p=4.3 x 10^-8 hypergeometric test). One of them, trancriptional repressor BACH1, engaged in anti-KSHV immune response (Botto et al., 2015), was already reported to be controlled by both miR-155 and miR-K11 (Gottwein et al., 2007; Skalsky et al., 2007).

In order to test regulatory potency of the discovered interactions involving viral miRNA, we subjected a subset of them to luciferase reporter assay. The reporter constructs harbored mutations only in miRNA binding sites on their 3'UTR's. Thus, the perturbation was localized to the binding sites rather than to the entire 3'UTRs. We started with a set of targets for miR-K11, as it was reported to be important for KSHV pathogenesis (Boss et al., 2011; Dahlke et al., 2012; Gottwein, 2012). This set included the genes with canonical seed matches (RORA, CLCN3) along with the genes harboring weaker recognition motifs (BCL2, STK38L, MYB, ZNF330, KHDRBS1, PUM2, YWHAZ). Here, we defined a binding site to be canonical if it has a match to miRNA seed 2-8 or 2-7a (convenient seed match plus an adenosine opposing the 1st nucleotide in miRNA). We observed a significant derepression upon binding site mutation for five out of nine miR-K11 targets (Fig. 9B), including non-canonical ones. Seed match to nucleotides 2-7 was enough to confer a potent regulation for BCL2 and STK38L, while MYB was repressed even via imperfect binding (Fig. 9D). Remarkably, exactly the genes downregulated by miR-K11 appeared to be valid targets for miR-155 (Fig. 9B). Therefore mimicking of the first eight nucleotides of the host miRNA seems to be enough for the virus to hijack its targets.

The role of the seed of miR-K11 for the tested interactions was emphasized by further disruption of the binding sites (Fig. 9C). Indeed, altering two nucleotides in the seed match region impaired miRNA regulation entirely.

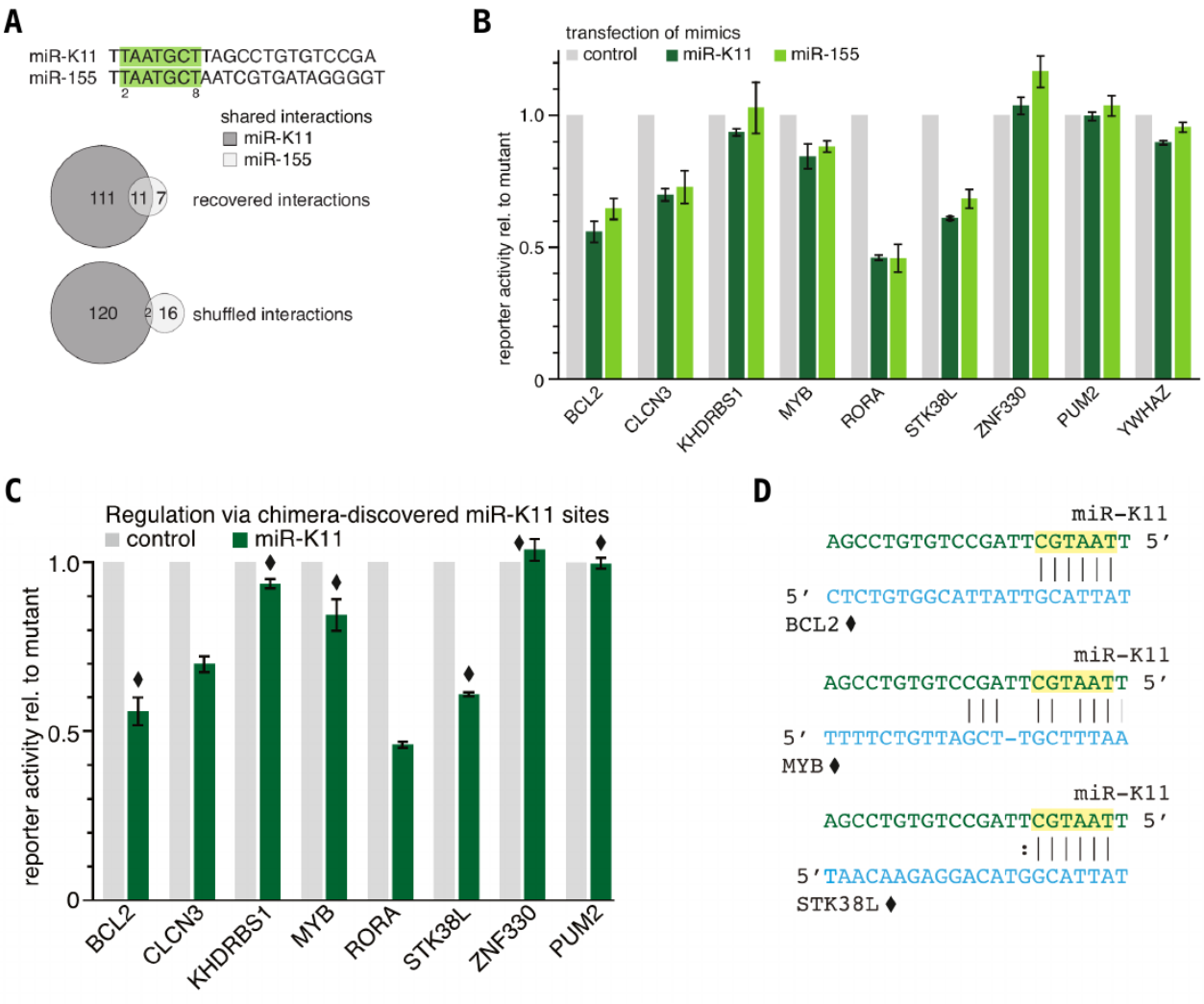

Figure 9| chimeras involving viral miRNA K11 represent functional interactions via non-conserved targets
(A) KSHV miR-K11 and human miR-155 are identical in sequence for nts 1–8; 11 common target sites were recovered via analysis of chimeras in datasets where either miR-K11 or miR-155 are expressed endogenously (Gottwein et al., 2011, Skalsky et al. 2012); this fraction of common sites is greater than expected by chance (shuffled interactions served as control; $p < 4.7 \times 10\text{-}25$).
(B) The eight miR-K11 binding sites tested in Figure 9C responded similarly to miR-155 as they did to miR-K11, suggesting that the identical seed region of these two miRNAs is critical where regulation was observed.
(C) The majority of tested, chimera-identified KSHV miR-K11 interactions resulted in specific reporter repression, including sites with weak seed matches. miRNA interactions were tested in dual luciferase reporter assays using WT and binding site mutant 3'UTR reporters and either control miRNA or viral miRNA mimics. Noncanonical interactions are marked with a diamond. Canonical, i.e., perfect match to miRNA position 2–7 with an A opposing the first miRNA nucleotide and/or perfect complementarity to at least miRNA positions 2–8; numbers are mean ± SEM ($n \geq 3$).
(D) Predicted base pairing for noncanonical miR-K11 sites that were responsive in the reporter assay.

As it was mentioned above, one of the main advantages of our method is the ability to confidently assign several miRNAs to a single binding site. In the case of virally infected cells this feature is of particular importance, since we are able to uncover the binding sites which are employed by both virus and its host. Thus, it is possible to track how virus utilizes already existing regulatory pathways. For instance, we found a binding site shared between viruses. It resides in the 3'UTR of YWHAZ and has chimeras with KSHV miR-K11 and EBV miR-BART14. Remarkably, these miRNAs are identical at nucleotides 3-7 (Fig. 10A), hence they can share some other targets. However, only miR-BART14 appeared to significantly repress the corresponding reporter construct within the sensitivity of our assay (Fig. 10B).

Figure 10| Two miRNAs from different viruses can share one binding site on a host transcript
(A) EBV miRNA BART-14 and KSHV miRNA-K11 are identical in only five positions (nt 3–7), but these might bind the same nucleotides in the target (noncanonical binding for miR-K11, canonical for miR-BART14).
(B) miR-BART14 appeared to significantly repress the reporter construct harboring YWHAZ 3'UTR. The repression exerted by miR-K11 occurred to be non-significant

Another mimic to a host miRNA is miR-K3 encoded by KSHV genome. It has been recently shown to interfere with human miR-23 through offset seed homology(Manzano et al., 2013). Like miR-23, miR-K3 lacks regulatory potency even for its canonical binding sites (Garcia et al., 2011; Manzano et al., 2013). Therefore we analyzed miR-K3 binding pattern based on chimera analysis. Surprisingly, we observed a relatively low pairing involving the seed site compare to other miRNAs (Fig. 11A). Indeed, only 7% of the discovered miR-K3 interactions have perfect 2-7 seed match. Furthermore, ~15% of miR-K3 interactions can be paired via so-called offset seed (miRNA positions 3-8), while only 6% of all other KSHV and human miRNA interactions use this binding mode. For the five tested noncanonical miR-K3 targets we did not observe a significant repression in luciferase reporter assay (Fig. 11B). Thus, miR-K3 emerges as a special case of miRNA with shifted seed site and low regulatory potency.

As sequence conservation is an important hallmark of potent regulatory binding sites, it is widely used in miRNA target predictions. This reasoning breaks in case of viral miRNAs with no seed homology to any conserved host miRNA. Indeed, there is no reason for human genome to keep a binding site for a viral agent. Our approach enables to discover non-conserved interactions between viral miRNAs and human transcripts, which is almost impossible for pure computational predictors.

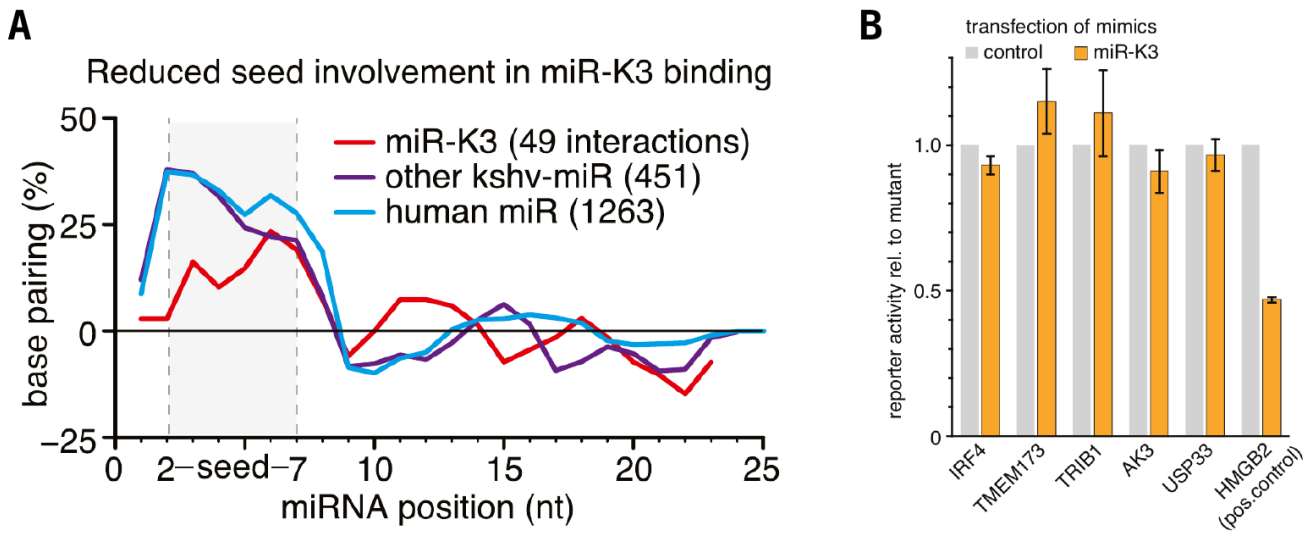

Figure 11| KSHV miRNA K3 recognizes its targets via shifted seed-complementary
(A) Hybridization profile of KSHV miR-K3 interactions compared to all other KSHV and human miRNA interactions identified by analyzing CLIP data by Gottwein et al. (2011) and Skalsky et al. (2012); miR-K3 interactions display reduced binding in the 5' region of the miRNA. RNAhybrid, G:U allowed, controls (permutation of dinucleotides in target sequences) were subtracted.
(B) 5 non-canonical chimera-identified miR-K3 interactions were tested in luciferase reporter assays. Repression was not detected for any of these sites, in line with the poor regulatory capacity previously reported for this miRNA in reporter assays.

4. Results (ChiFlex)

4.1 ChiFlex is a tool to discover miRNA:target interactions

Even though we were satisfied with the obtained results, we set out to improve the performance of the computational pipeline and extend its functionality. First, we decided to compile all the scripts used for the chimera discovery into one computational tool, called ChiFlex. In contrast to the original design, ChiFlex does not require manual curation and assigns the filtering thresholds automatically. Thus, the analysis can be run with one single command, which makes ChiFlex distributable and easy-to-use for other researchers. We also improved the speed of our pipeline using Bowtie2 for the mapping to miRNA reference and implemented parallel computation at time-consuming steps. Moreover, we granted ChiFlex an ability to find any type of chimeric read, while previously we were constrained to miRNA:target chimeras. Finally, we developed the filtering strategy to keep the actual false discovery rate below the predefined one, while being as sensitive as possible. Thus, ChiFlex is a standalone tool to discover any type of RNA:RNA chimeras with a controlled specificity. Here we provide a condensed overview of the most important aspects of our method (see methods for the details).

Comparing to conventional mapping, chimera detection is more vulnerable to the uninformative nucleotides inside the reads. Therefore ChiFlex benefits from the thorough preprocessing of raw sequencing data. Excision of the 5' adapter or barcode is particularly influential on miRNA:target identification, since read alignment to a miRNA starting from the first nucleotide is a decent hallmark of a true positive chimera. Removal of the 3'end adapters and/or barcodes is also important, since it reduces the search space for the potential miRNA target. The same reasoning holds true for the nucleotides sequenced with low quality, as they tend to reside at the 3'ends of the reads. Furthermore, for several AGO-CLIP datasets we found around 70% of the reads being exclusively adapter concatemers, and filtering them out improved both sensitivity and specificity. Therefore, extensive *in silico* reads' purification is an important prerequisite for chimera discovery. Typically, we use a combination of FastQC (Andrews, 2010) to explore potential contaminations in the sequencing reads and Flexbar (Dodt et al., 2012) to successively remove them.

ChiFlex starts with the mapping of already preprocessed reads to a set of miRNAs (Fig. 12). Typically all the miRNAs for given specie are indexed with Bowtie2 into a 'miRNA reference', a specifically encrypted object allowing fast mapping to the underlying sequences. Since miRNAs' sequences are relatively short, it is possible to confidently assign a read to a miRNA with perfect alignment of only 12 nucleotides. Despite miRNA:target chimeras generated by AGO-CLIP experiments are supposed to start with a miRNA sequence, ChiFlex looks for the valid alignments throughout the whole read. Therefore, our method is flexible to adjust to various experimental setups and overcomes the problem

of incomplete adapter removal. However, the start position of the alignment to miRNA on a read is still an indication of chimera quality, and is used as a parameter at the filtering step.

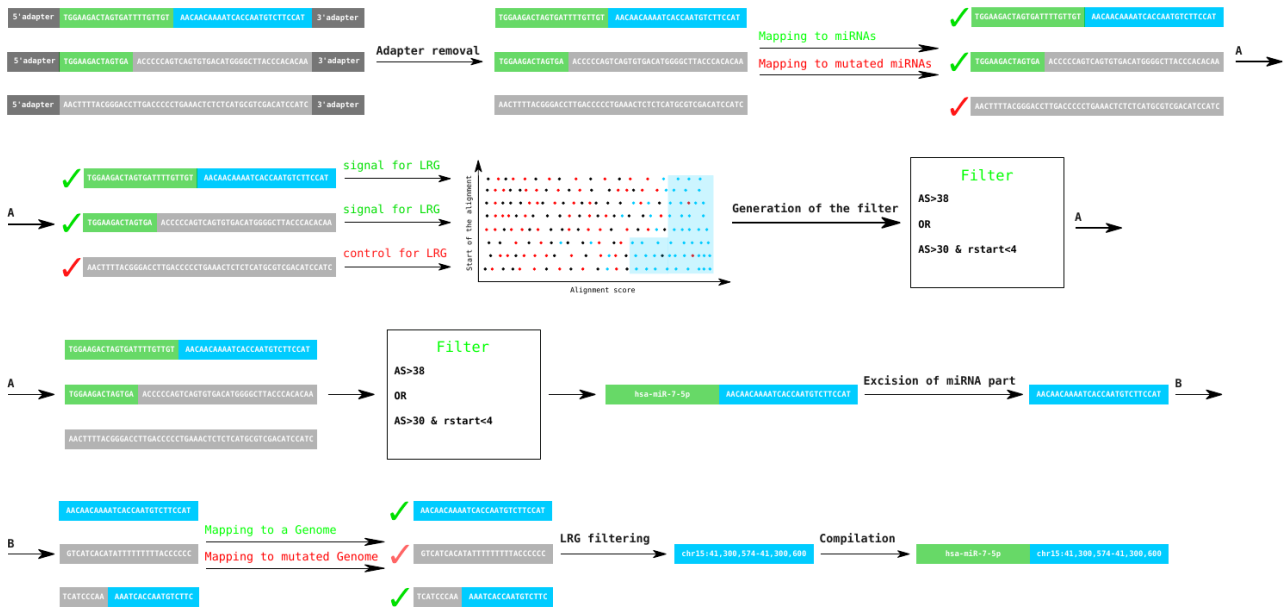

Figure 12| Overview of ChiFlex workflow
ChiFlex benefits from running on purified reads, with adapters and barcodes already removed. However preprocessing is not automated and included in the package so far. ChiFlex starts with the mapping to miRNAs and to miRNAs with introduced mutations ('true' and 'control references). Reads mapped to the 'true' reference are separated from reads mapped to 'control' by LRG. LRG finds the thresholds based on mapping properties to achieve suboptimal sensitivity of filtering, while keeping false discovery rate lower than the desired value. As attributes for the discriminative rules, LRG uses: alignment score of the mapping, start position of the alignment on a miRNA and start position of the alignment on a read. Reads fulfilling the rules generated by LRG are then mapped to the potential targetome reference (genome, transcriptome, 3'UTRs and etc.) and to its mutated counterpart. In order to reduce the search space of the alignment to the reference we excised out miRNA part from the read prior the mapping. Reads mapped to the 'true' reference are filtered with LRG in the way described for mapping to the miRNAs. The only difference is that the start of the alignment on a reference is not used as an attribute in the filtering rules. Finally, chimeric reads are compiled into miRNA:target interactions.

The filtering is a paramount part of the whole algorithm. At this step ChiFlex selects valid mappings keeping false discovery rate below the predefined threshold, typically 5%. In order to estimate FDR we have first to understand what the false positive mappings are. A mapper (Bowtie2 in our case) reports for each read all potential alignments with a score above a particular cutoff. That is, each read is assigned to a region on a mapping reference, or to multiple regions. Some of these read:region pairs are correct, which means that the read indeed originates from the region. Some are incorrect, which means that read has another origin than reported. Let's imagine that we mapped all the reads to randomly generated sequences (Fig. 13). Since randomly generated sequences cannot be the origins of the reads, all the reported mappings are false positives. If the properties of these false mappings to the random reference are globally similar to those of false mappings to the 'real' reference, it is possible to replace the latter with the former in the false discovery rate estimation. The only difference is *a priori* knowledge, that mappings to a random reference are wrong. The prerequisite of global properties being similar for false mappings to random and real references is crucial for FDR estimation. Indeed, if a randomly-generated reference is 10 times shorter than the real one, then mappings to the former will in general have lower alignment scores than to the latter. Consequently, more mappings to the real reference pass the alignment score cutoff than mappings to the random reference pass exactly the same threshold. Therefore, FDR estimation based on such a control reference will not reflect the actual false discovery rate. Thus, control reference has to be the same size as the original one. The same holds true for local nucleotide composition. For example, reads will not be eagerly mapped to the long stretches of cytosines. It turned out (see 4.2) that introducing random mutations with 15-20% probability to the original reference generates a control, which emulates false mappings with a decent precision.

We denied using rigid cutoffs for the mappings, and then reporting false discovery rate. In opposite, we first set desired FDR and then adjust the thresholds correspondingly. The rationale is to allow for specificity control for the wide range of datasets. For example, FDR will be lower than desired value for 100 million reads of 50nt length and for 10000 thousand reads of 100nt, regardless of the dataset’s peculiarities. Selecting the set of thresholds to ensure low FDR is a trivial task - one can simply set the most stringent cutoffs. In opposite, maximizing sensitivity while keeping high specificity is a difficult exercise we address here with Logic Rule Generator (LRG).

LRG looks for a combination of filtering rules, which allows passing as many mappings as possible to the real reference, while keeping FDR below the predefined cutoff. For example, 1.000 real mappings and 20 control pass the rule: "alignment score" > 34 and "start of the alignment on the read" <= 1. Then, this rule can be used for the filtering with reported 2% FDR. The thresholds for this rule were iteratively expanded to achieve suboptimal sensitivity (see methods). However LRG guaranties that further relaxation of the cutoffs will not improve the performance. Let's demand

"alignment score" > 32 and "start of the alignment on the read" <= 1. Then, for example, 2000 real mappings and 80 controls pass the rule. FDR is still below desired 5% and we got two times more reads. However, if one checks the additional gain, there will be 1000 real mappings and 60 controls, which does not comply the preset false discovery rate. LRG will not allow these additional mappings to pass, since we want to be confident in estimated FDR and not to enter regions of marginal specificity. Another reason is that we want to ensure that subsamples of the filtered mappings will have a decent specificity. For example, miR-2017 was found in 100 reads, all with the alignment score less than 34. Then all further inferences for this miRNA are not valid, since for these alignment scores (<34) FDR is unacceptably high.

In order to overcome an overfitting problem only the rules supported by a considerable number of unique reads are selected. That is, if 1000 out of 10000 reads follow the rule 'alignment score > 42' and only this rule, then it is used at the filtering step. LRG tries to find all the rules with a decent read support and low FDR. Finally the union of all valid rules is applied for the filtering. Remarkably, LRG can be used to discriminate noise from signal with a specific focus on FDR for a variety of cases, and is not constrained to the bioinformatics. Thus, Logic Rule Generator is both an integral part of ChiFlex and a standalone filtering tool.

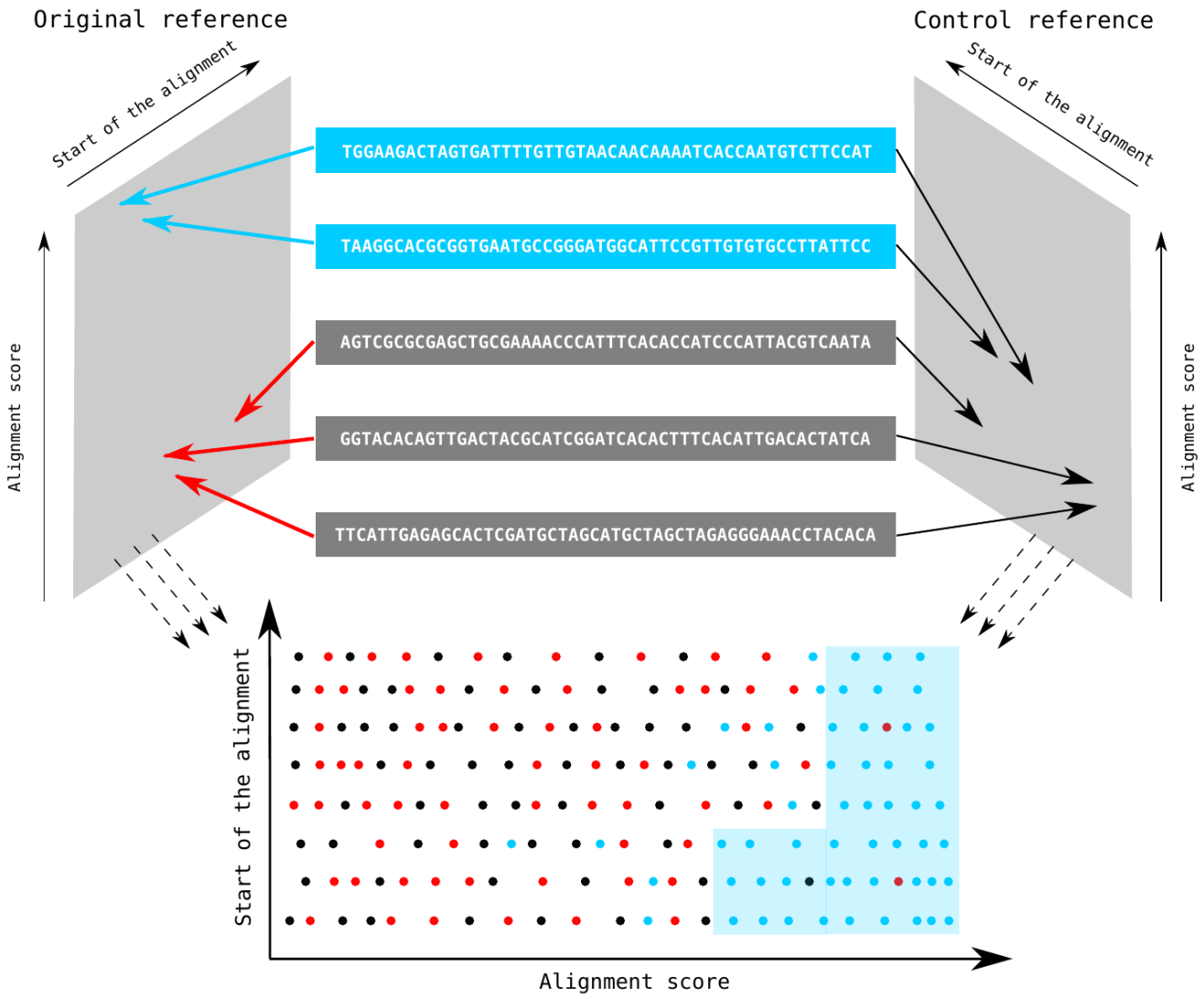

Figure 13| Overview of the filtering strategy

ChiFlex estimates false discovery rate via mapping to a control reference. The underlying assumption is that properties of the mappings to the control reference are generally the same as those of the erroneous mappings to the original reference. Further, based on the control mappings, the Logic Rule Generator adjusts the filtering thresholds in a way that ensures the desired specificity along with suboptimal sensitivity.

The parts of the filtered reads adjacent to miRNA segments are then extracted and mapped to a reference of potential binding sites: genome, transcriptome, 3'UTR's, etc. As it was described above, they are also mapped to a 'mutated' reference, giving rise to the control mappings. These controls are used for the second round of filtering, which also employs LRG. Thus, ChiFlex ensures that the probability to find either erroneously detected miRNA or target part is less than the preset FDR value.

To be even more quality-demanding we apply one additional filter on chimeras' architecture. As chimeric reads are byproducts of ligation reaction, one can expect no gap between miRNA and target parts. Since the probability for N nucleotides being shared between miRNA and target part is $(1/4)^N$, a large overlap is also unlikely. Therefore ChiFlex does not allow any gap and an overlap more than 4 nucleotides.

Finally we compiled all the chimeras originated from the same miRNA and target region into a single miRNA:target interaction. Thus, ChiFlex outputs a compendium of reliable miRNA:target pairs, unambiguously detected in the provided dataset. Furthermore, ChiFlex is augmented with a set of standalone tools for downstream analysis of miRNA:target interactions, including search for well-known miRNA binding modes and *in silico* hybridization of miRNA:target duplex.

Remarkably, ChiFlex has two optional improvements. According to a conventional bioinformatics methodology non-uniquely mapped reads are discarded. Since miRNAs from the same family (e.g. let-7a, let-7b) share up to first 12-18 nucleotides, many chimeras may be filtered out because of the ambiguous mapping to a miRNA reference. Moreover, the mappings to the binding sites duplicated in a genome may be also lost. Therefore ChiFlex optionally selects only one mapping for a particular read, while others are stored in a separate file. Further the information on the ambiguous mappings can be restored to reassign the number of reads supporting miRNA:target interactions (see methods). Another improvement comes from the empirical observation, that discarding repetitive sequences boosts both sensitivity and specificity. In contrast to the conventional approach of masking repetitive regions on a reference, ChiFlex evaluates Shanon entropy for each aligned sequence and discards those with a low score. Thus, ChiFlex is able to remove mappings directly based on their low information content, regardless external annotations.

4.2 ChiFlex performs with a controlled specificity

ChiFlex was primarily designed to discover miRNA:target pairs. It can be also used to find interactions involving any small RNA and any RNA molecule, for example piRNA:target pairs. However, one of the interactors must be a small RNA. We set out to overcome this constrain, and grant ChiFlex an ability to find any type of RNA:RNA chimera. In contrast to miRNA:target, RNA:RNA chimeras may arise from a number of sources. For example, a read covering linear or circular splice junction is considered as chimeric, since it cannot be mapped continuously to a genome. Chimeras may also arise from the exotic events like gene fusions and trans-splicing. Even though ChiFlex has an option to discover alternative splicing or circular RNAs, we set out to not compete with the state-of-art tools. Instead we focused on RNA:RNA chimeras arising from the direct physical interactions as it is for miRNA:target case. The discovery of RNA:RNA interactions may be considered as more complicated

than analysis of alternative/circular splicing, since one cannot rely anymore on the presence of the canonical splicing signal and genomic distance constrains on the mappings of chimeras' parts. On the other hand, greater flexibility is achieved at the cost of performance; hence ChiFlex may be inferior in splicing detection to the tools specifically designed for this purpose.

The search for RNA:RNA chimeras required several major changes in ChiFlex workflow. Since we cannot assume anymore that one part of a chimeric read originates from a particular set of transcripts (miRNAs), ChiFlex performs only one round of mapping, typically to a genome or a transcriptome. Hence, we cannot compile chimeras successively as for miRNA:RNA case. Instead we have to check all possible combinations of the alignments for a particular read to find those which may constitute a valid chimera. In the worst scenario one single read can be explained by multiple chimeras and multiple continuous mappings. ChiFlex addresses these problems via scoring all potential alignments and their combinations. If there is one unanimous winner and its score is substantially higher than for the others, then this continuous mapping or chimera is selected for the following filtering. As it was for miRNA:RNA, ChiFlex maps the reads to a control reference. The chimeras with at least one part mapped to the control reference, are considered as known false positives. LRG then discriminates signal from control in the way described above, but taking mapping properties of two chimeric parts simultaneously.

Since the discovery of RNA:RNA chimeras is a non-trivial task, we set out for an extensive testing of ChiFlex performance. Remarkably, if ChiFlex was proved to operate well for RNA:RNA case, then its decent performance for miRNA:RNA case is ensured, as the latter is the subproblem of the former. Therefore we generated artificial chimeras and tested whether ChiFlex was able to find them. Briefly: two sequences were randomly extracted from exons, introns or intergenic regions of human genome (hg38 genome assembly). They were further concatenated into chimeric reads. The length of those simulated chimeras was set to 50, 75, 100 or 150 nucleotides which correspond to the insert lengths of typical RNA sequencing protocols. Then, random mutations were introduced into the chimeric sequences at rate of 0, 1, 3 or 5% reflecting sequencing errors and biological variability. Finally, continuous sequences, extracted from human genome, were added to mimic real-life situation, where chimeras represent only a minor fraction of sequencing data. The combined reads were then passed to ChiFlex. Since we *a priori* knew the origins of the reads, we were able to compare them with those reported by ChiFlex. Furthermore, in this testing setup we could track the evolution of specificity and sensitivity at each step of our pipeline.

For the *in silico* generated set of 100nt reads with 3% of mutation frequency ChiFlex was able to recover 43.0% of RNA:RNA chimeras (Fig. 14A). Closer inspection of the false positives displayed that the vast majority of the retracted chimeras were interpreted as continuous mappings. That is, one of the chimeric parts was mapped correctly, while the other could not be confidently detected because

of short length or/and high number of mutations. Thus, ChiFlex was able to recover some useful information even from the retracted chimeras. We also tested ChiFlex specificity on the same dataset. Prior the filtering step 17.2% of reported chimeras appeared to be false positives (Fig. 14B). The majority of these erroneously assigned chimeras (Fig. 14B) were indeed conventional reads, which had to be mapped continuously. Most probably, some parts of these sequences became more similar to the other genomic regions because of the introduced mutations, and consequently were mapped to the wrong loci. This explanation is supported by a clear enrichment of the reads originated from intergenic and intronic regions among the false positives (Fig. 14C), since they have less sequence heterogeneity than the exonic ones. As we expected, the majority of false positives possessed the characteristics generally different from the genuine chimeras, and were consequently filtered out (Fig. 14D).

Previously we discussed that the control reference used for the filtering should emulate the origins of the false discoveries as precisely as possible. With the *in silico* generated reads we could compare the real and the reported FDR. As the convergence of the real and the reported FDR is a hallmark of a proper filtering, we were able to test different schemes to set up a control reference. We selected two basic strategies: shuffling of the nucleotides in the original reference or mutating them with a moderate frequency (15%). Remarkably, the 'shuffling' scheme was clearly outperformed by the 'mutating' approach, especially for the low quality reads (Fig. 15). More important, the reported FDR was consistently above the real one. Thus, ChiFlex proves to recover chimeras with a trustworthy false discovery rate even from the datasets heavily affected with sequencing errors and single nucleotide polymorphisms.

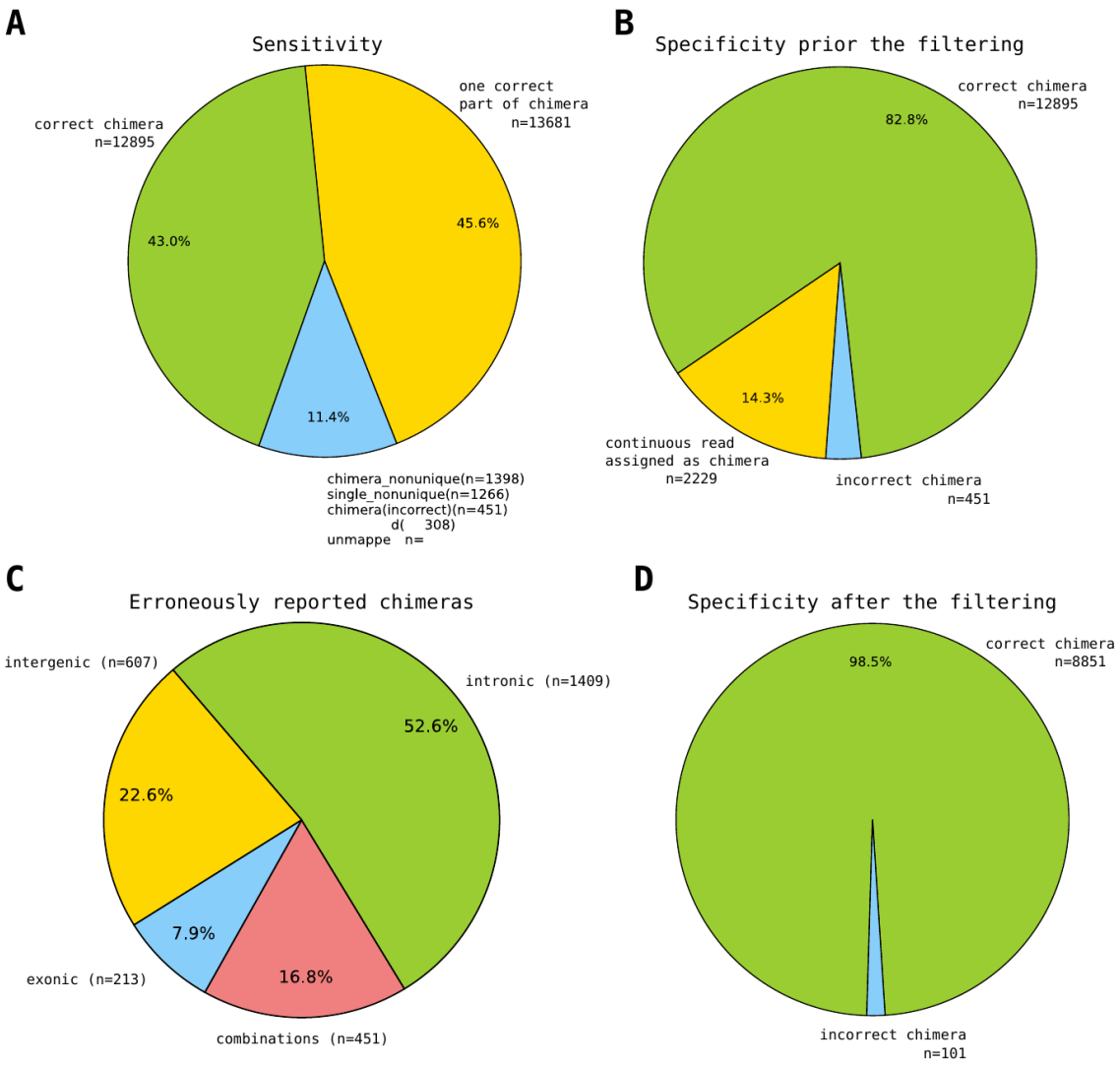

Figure 14| Testing of ChiFlex performance in RNA:RNA chimeras discovery
(A) To test ChiFlex sensitivity and specificity, sequences were randomly selected from exonic regions and concatenated into 100 nucleotide long chimeric reads. Each nucleotide of the chimeras was mutated with 3% probability. The resulting sequences were passed to ChiFlex. For 43.0% of the chimeras both parts were correctly annotated, while for the rest only one part was detected unambiguously. 'correct chimera' stands for correctly and unambiguously identified chimeric reads, 'one correct part of chimera' stands for the case when only one part of chimera was identified correctly; 'chimera_nonunique' stands for chimeras rejected by ChiFlex as being mapped ambiguously; 'single_nonunique' stands for the case when one part of chimera was identified ambiguously; 'chimera(incorrect)' stands for erroneously mapped chimeric read; 'unmapped' stands for chimeras which fail to be mapped even partially.
(B) Among the reads declared by ChiFlex as valid chimeras only 82.8% were indeed true ones. The main sources of false positives were misannotated single reads and chimeras mapped to wrong loci. Specificity of chimera detection is shown prior the filtering. 'correct chimera' stands for the correctly and unambiguously identified chimeric reads, 'continuous read assigned as chimera' stands for the conventional read misannotated as chimera, 'incorrect chimera' stands for a case when a read was identified as chimera but at least one part was mapped to the wrong locus. The tested reads are the same as in (A).
(C) Sources of false positives in the chimera discovery. The majority of the conventional continuous reads misannotated as chimeras are originated from intronic and intergenic regions. 'Combinations' stands for the erroneously mapped chimeras composed of sequences with different origins.
(D) The filtering dramatically reduces false discovery rate of chimeric reads. However, 31% true positives were lost in the filtering step. The labels are the same as in (B). The tested reads are the same as in (A) and (B).

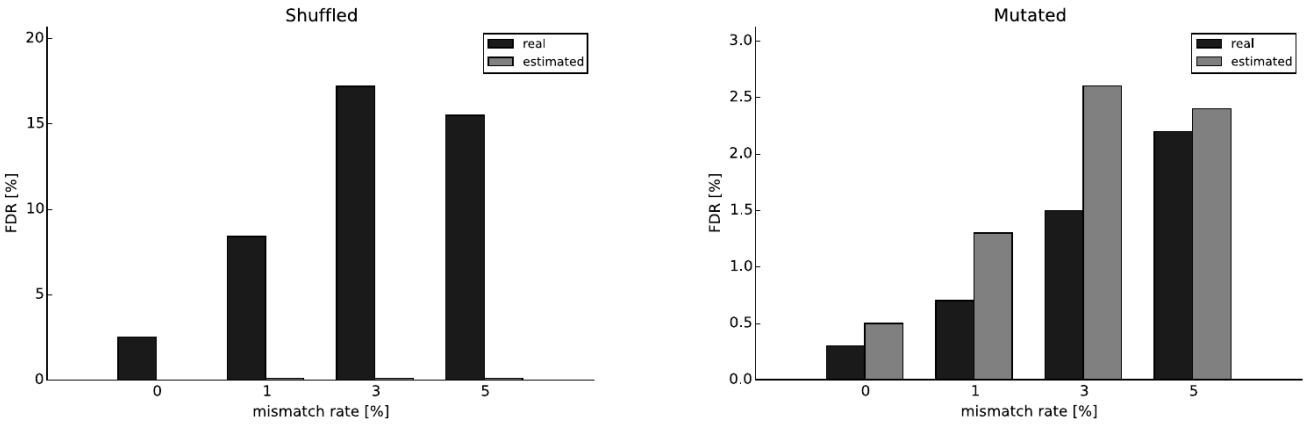

Figure 15| Comparison of the background models used for the filtering
Since the custom filtering method (LRG) reports false discovery rate (light gray), we can compare it to the empirical FDR based on simulated chimeras (dark gray). The closer the estimation to the true value, the better the filtering. The assessment of the filtering performance was done for 100nt chimeras with the mutation rate increasing from 0 to 5%. We used two different approaches to generate control reference. Figure (A) corresponds to the shuffled human genome sequences, keeping dinucleotide composition unchanged. Figure (B) corresponds to the human genome sequences mutated with 15% probability

4.3 Exploration of miRNA targeting in human brain

A few new AGO-CLIP datasets were submitted to a public access since we started the work on ChiFlex development. We decided to grab an opportunity to test ChiFlex performance in real-life environment and explore miRNA targeting in the corresponding biological systems. We started with the analysis of the experiment performed by Boudreau and colleagues (Boudreau et al., 2014). The authors collected 11 postmortem brain samples (one sample for one person) and passed them through AGO HITS-CLIP protocol. We downloaded the raw sequencing data and preprocessed them according to the authors' guideline. The purified reads were then successively mapped to the set of miRNAs (MirBase version21) and human genome (hg38 genome assembly). We applied three additional filters besides the standard LRG filter. First, we removed the mappings to repetitive sequences. Second, we discarded all the chimeras involving ribosomal RNAs. Third, we filtered out the interactions between mature and star miRNAs. In total around 100 000 chimeric reads passed these criteria, giving rise to more than 23 000 unique miRNA:target pairs (Fig. 16A). The estimated FDR for the discovered interactions was lower than 3%.

The recovered interactions generally followed miRNA binding rules supporting the credibility of ChiFlex. However, the usage of pairing via seed region was reduced in brain compare to HEK cells (40% vs 75%) (Fig. 16D). The loss in binding via seed might be compensated via increased pairing to the miRNA 3'end (Fig. 16E), which was also reported for mouse brain (Moore et al., 2015). However,

we cannot directly assign miRNA targeting via 3'end to be specific for neuronal tissue. Indeed, chimeric reads are products of a ligation reaction between miRNA 3'end and target 5'end. Hence, the interactions, having these tails already bound to each other, may be favored because of the spatial arrangement suitable for the ligation. Thus, the enhanced 3'end binding might be caused by an experimental bias, and the strength of this bias depends on a particular protocol.

As it was mentioned in the introduction, miRNAs typically target protein coding transcripts via binding their 3'UTRs. Indeed, we found the majority of miRNA binding sites located on protein coding genes (Fig. 16B). However, the binding was skewed towards coding regions, rather than 3'UTRs (Fig. 16C). This unexpected targeting may arise from a great number of weak and transient miRNA:CDS interactions, which are supported by small numbers of chimeric reads. In contrast to this assumption, the mean number of chimeras supporting miRNA:3'UTR interactions is even slightly less than for the interactions involving CDS regions (15.8% vs 19.4%). Then, we assumed that the discovered miRNA:CDS interactions are indeed miRNA:circRNAs, since circular RNAs are specifically abundant in mammalian brain (Rybak-Wolf et al., 2015) and some of them were shown to interact with miRNAs (Hansen et al., 2011; Memczak et al., 2013). To test this hypothesis we contrasted miRNA:targets with circular RNAs discovered by Rybak and colleagues in human brain (Rybak-Wolf et al., 2015). We found 35% of binding sites on CDS overlapping the brain-expressed circRNAs, while for the 3'UTRs this fraction was around 19%. However this difference cannot fully explain the miRNA tendency to bind CDS in human brain.

To assess the quality of our discoveries from another point of view, we tested if they are supported by state-of-art bioinformatics tools. For this purpose we selected one of the most popular tool for miRNA binding sites prediction: TargetScan (Lewis et al., 2005). We collected all miRNA:target pairs from the latest TargetScan release (version 7.1) and compared them to the interactions discovered in human brain. As TargetScan requires a seed-match to predict the binding sites, we constrained the comparison to the miRNA:targets paired via the seed region. Only 10% of the interactions recovered by ChiFlex were found in TargetScan predictions. On one hand, it supports the fidelity of ChiFlex discoveries, since this overlap is greater than expected for the shuffled miRNA:target pairs (less than 1%). On the other hand, it means that a big fraction of miRNA:target interactions found by ChiFlex cannot be predicted by pure computational efforts.

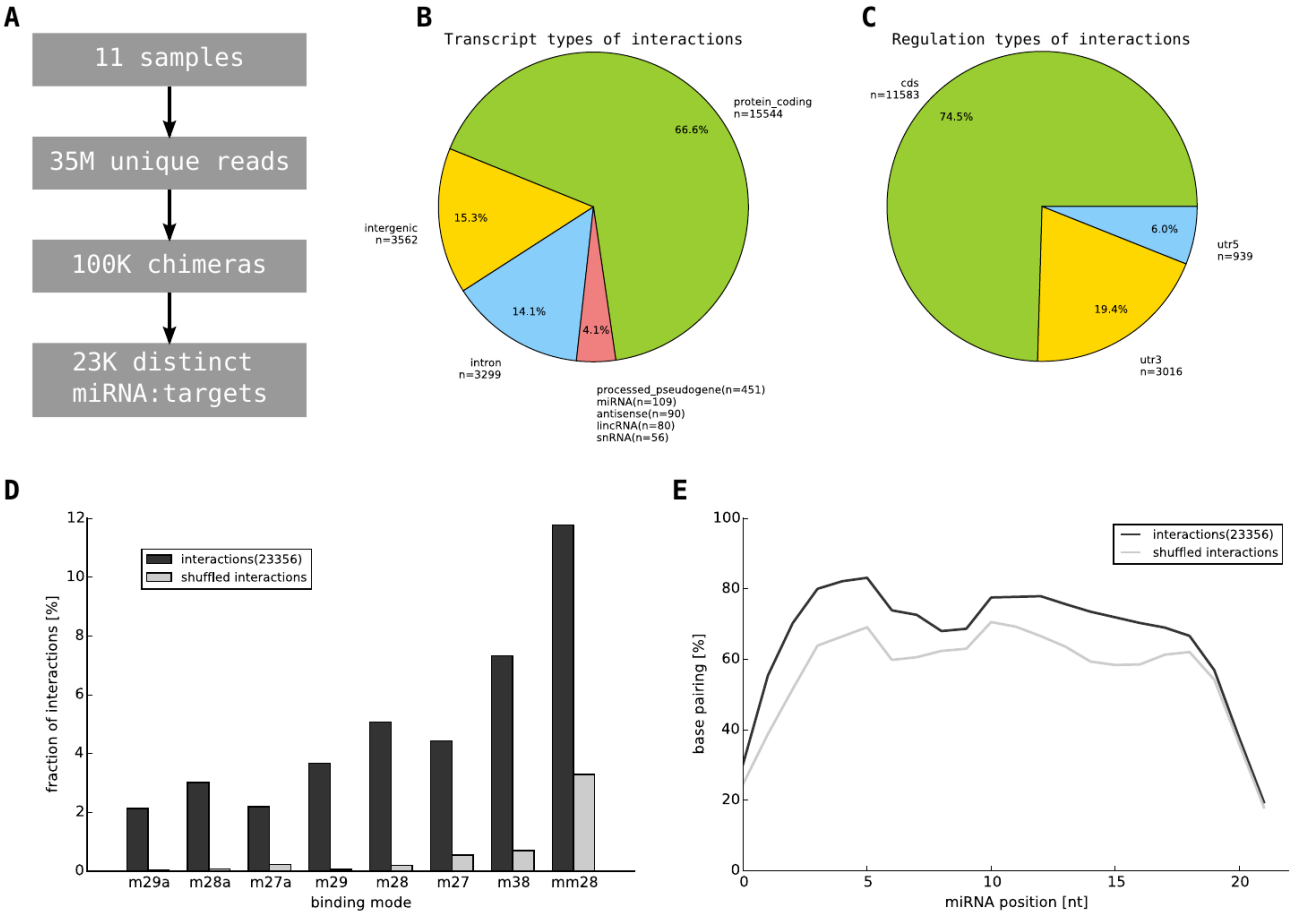

Figure 16| Discovery of the miRNA:target interactions in human brain

(A) Reads from eleven samples, each corresponding to an individual, were merged and processed according to the original guidelines. More than 10^5 miRNA:target chimeras were identified out of $3.5x10^6$ unique reads, giving rise to 23,256 unambiguous miRNA:target pairs.
(B) We intersected genomic coordinates of miRNA binding sites with those of transcripts annotated in ENSEMBL v.84. If a miRNA target was found to overlap more than one genomic feature, its count was split equally between features. Further, counts were summarized by transcript type.
(C) For the protein-coding transcripts, counts were accumulated according to translational unit type: CDS, 5'UTR, 3'UTR. As in (B), counts were split equally between CDS, 5'UTR and 3'UTR in case of multiple overlap.
(D) We tested if the discovered miRNA:targets interact via previously described binding modes: matches to nucleotides 2-7, 2-8 and 2-9 with (m27a, m28a, m29) or without (m27, m28, m29) adenosine opposite to the first nucleotide on miRNA, match to nucleotide 3-8 (m38), match to nucleotide 2-8 with one mismatch allowed (mm28). To check the significance of the enriched binding modes among miRNA:targets we shuffled these pairs one hundred times, and looked for seed matches among the permuted interactions for comparison.
(E) miRNA:target pairs were *in silico* hybridized using RNAhybrid (Krueger et al., 2006) with default settings to generate a hybridization profile. The hybridization profile shows a fraction of miRNA:target interactions with a nucleotide paired at a particular position on miRNA. As expected, we observed elevated hybridization for the seed region. Remarkably, we also detected a strong enrichment for the miRNA's 3' end binding, which may be specific for miRNA function in brain or caused by experimental biases. Similar to (D), background hybridization profile was generated via shuffling of miRNA:target pairs.

Thus, miRNA:target interactions comply with the well-established rules of miRNA binding and are supported by an independent computational method. However, in order to explore not only qualitative but also quantitative aspects of miRNA targeting, the numbers of chimeras must correlate with the expression levels of mRNAs and miRNAs involved in miRNA:target interactions. As the counts of genes occurred in miRNA:target interactions are generally low, only the comparison for miRNAs was performed. We developed a tool to quantify miRNAs expression (see methods) and applied it to the small RNA sequencing data derived from the same human brain samples. For the comparison we took only the well-expressed miRNAs (>0.1% of total miRNA pool), since miRNAs with low expression may not form even a single chimera. We found rather high correlation between miRNA expression and their occurrence in chimeras (Fig. 17, Pearson correlation coefficient 0.74, Spearman rank correlation 0.60). Thus we can assume that the numbers of chimeric reads supporting miRNA:target interactions are roughly proportional to the expression levels of the corresponding miRNAs.

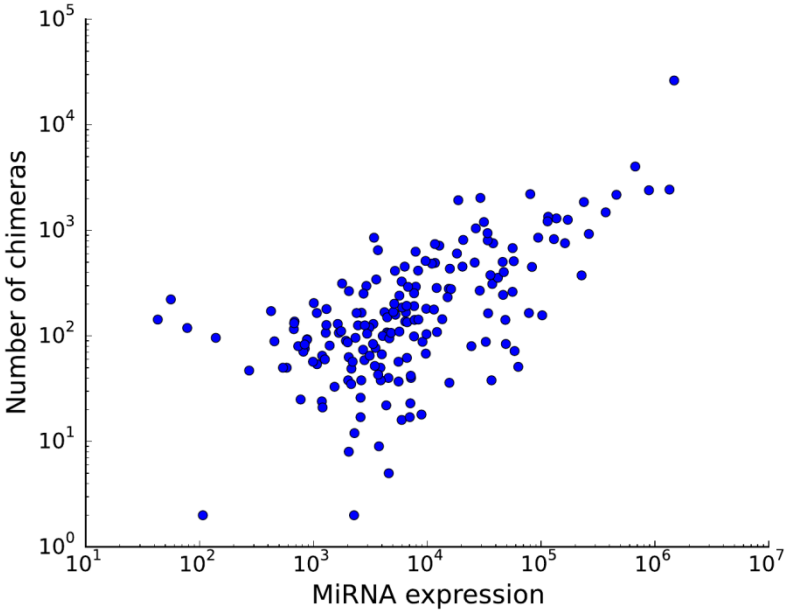

Figure 17| Numbers of miRNA:target chimeras correlate with miRNA expression

We counted the expression levels of human miRNAs via analysis of small RNA sequencing data derived for the same samples as human brain Ago-CLIP data. We contrasted these expression values to the miRNA's occurrences in chimeric reads. The probability of a miRNA to be involved in miRNA:target appeared to be roughly proportional to its expression (Pearson correlation coefficient 0.74, Spearman rank correlation 0.60).

4.4 The analysis of chimeras revealed regulatory modules in mammalian brain

We set out to explore miRNA functionality based on the discovered miRNA:target interactions. We started with a global analysis of biological processes attributed to miRNAs. For each miRNA we compiled a list of the targeted genes and subjected it to the GO term enrichment analysis (Ashburner et al., 2000). Since the analyzed dataset came from human brain, it was expected that miRNAs targets tend to have neuron-related functionality. In order to avoid these trivial inferences we used the list of all genes found in miRNA:target interactions as a superset for GO analysis, rather than a list of all human genes. Thus, we looked for the biological functions controlled by a particular miRNA, but not by miRNAs in general. In total we discovered 84 unique associations between a miRNA and a biological function with a high confidence ($p < 0.01$ hypergeometric test adjusted for multiple testing). Remarkably, a number of miRNAs occurred to control functions which cannot be directly addressed to neurons. In contrast, miR-124 appeared to specifically regulate neuron-related processes compare to the other miRNAs (Fig. 18B, 9 out of 10 GO categories, while 16 out 31 for all miRNAs; $p < 0.005$ hypergeometric test).

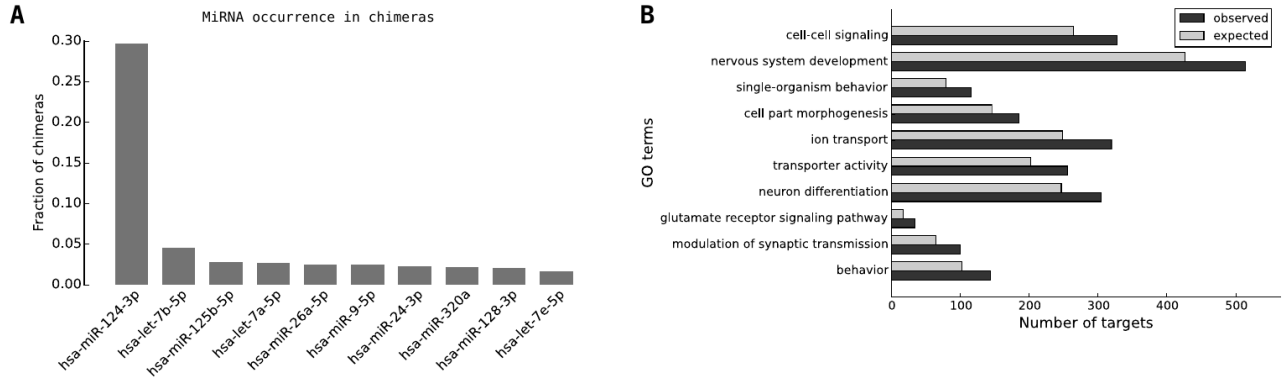

Figure 18| MiR-124 is abundant in human brain and specifically regulates neuron-related pathways

(A) The targets of miR-124 were significantly enriched in the depicted GO terms. The 'observed' value stands for the number of miR-124 targets supporting a GO term, the 'expected' value stands for the most probable number of targets for the same GO term according to the hypergeometric distribution.
(B) For each miRNA we calculated the number of chimeric reads involving it. These counts were normalized to the total number of chimeras. Thus, we got miRNA distribution among discovered miRNA:RNA chimeras. This distribution follows a typical miRNA expression pattern, with one being sharply on top and dozens of others well-expressed

MiR-124 constitutes around 30% of total miRNA pool in human brain (Fig. 18A) and participates in the most abundant miRNA:gene interactions found in human brain. However, if we set miR-124 apart, top (in terms of numbers of chimeras) miRNA:gene pairs are distributed quite equally among the other well-expressed miRNAs (Fig. 19A). One of these top interactions was particularly compelling. CDR1-AS is a circular RNA, that was shown to be a functionally relevant sponge for miR-7 (Memczak et al., 2013), a miRNA proved to be important for brain function and development (de Chevigny et al., 2012). That is, CDR1-AS sequesters miR-7 via multiple binding sites from the pool of potential Argonaute guides. Furthermore, miR-7:CDR1-AS interaction is the top-discovered association between a miRNA and non-coding RNA in human brain based on a number of chimeras (Fig. 19A). Finally, CDR1-AS is almost the exclusive target for miR-7 (Fig. 19B) and densely covered by miR-7 binding sites (Fig. 19C), which corresponds its 'sponging' function.

Being a functionally important target for miR-7, CDR1-AS also participates in the interactions with other miRNAs. It harbors an extraordinary binding site for miR-671. Indeed, in terms of hybridization energy miR-671:CDR1-AS interaction is the strongest among all the interactions discovered in human brain. MiR-671 and its binding site on CDR1-AS hybridize almost perfectly, with just one bulge. Consequently, miR-671 can act as siRNA, slicing the targeted molecule into pieces and destroy CDR1-AS via direct cut. Indeed, Hansen and colleagues reported that expression levels of miR-671 and CDR1-AS are anticorrelated (Hansen et al., 2011). Being a circular RNA, CDR1-AS is resistant to exonucleases and cannot be degraded via conventional degradation pathways. Therefore, the interaction with miR-671 may play a crucial role in CDR1-AS turnover.

We then decided to check if our findings regarding CDR1-AS can be reproduced in other species. Fortunately, Moore and colleagues performed AGO-CLIP experiments in mouse brain with an additional ligation step (coined as CLEAR protocol) to generate miRNA:target chimeras. The authors reported around 130.000 miRNA:target interactions. We ran ChiFlex on their dataset and discovered a comparable number of interactions (190.000). Since we captured the same general trends of extensive 3'end binding and a diminished pairing via seed region as Moore et al., we decided not to explore these interactions further. Instead, we specifically focused on the miRNAs targeting CDR1-AS. Notably, we observed the same regulatory patterns as for human brain (Fig. 19C). The CDR1-AS was covered by multiple binding sites for miR-7, while the highest binding peak on the transcript was for almost fully complementary miR-671.

CDR1-AS was shown to sequester miR-7 from a pool of available miRNAs. Therefore in order to find which genes are regulated by CDR1-AS, we set out to explore other miR-7 targets. Intriguingly, the second top-discovered target for miR-7 occurred to be another non-coding RNA, called Cyrano. This gene has an outstandingly (for non coding RNAs) conserved region around 300 nts long, with a binding site for miR-7 right in the middle of this conserved locus (Fig 19D). Moreover, miR-7 targets

Cyrano via a strong complementarity, with only the 10^{th} and 11^{th} nucleotides being unpaired (Fig. 19E). This hybridization pattern (full complementarity with a central bulge) was previously reported to cause miRNA degradation (de la Mata et al., 2015). That is, Cyrano has a potency to destroy miR-7 and this functionality is well-conserved among a variety of species. Thus, Cyrano is a part of complex regulatory module involving CDR1-AS, miR-7 and miR-671.

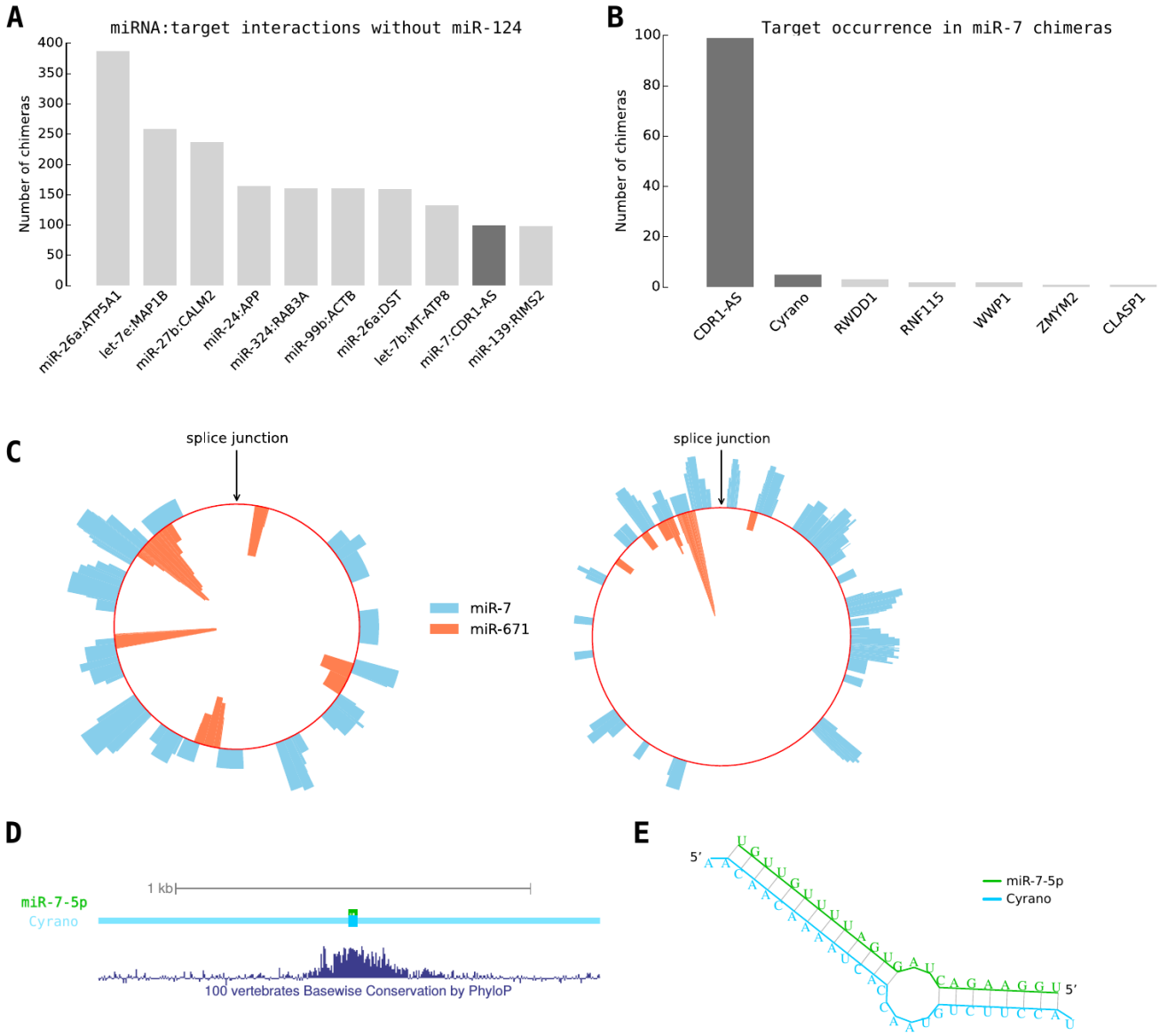

Figure 19| MiR-7, miR-124, CDR1-AS and Cyrano may constitute a regulatory interplay in mammalian brain

(A) miRNA:gene interactions were sorted according to the number of chimeric reads supporting them. The interactions with miR-124 were excluded. The miRNA:gene pair involving a non-coding target transcript is highlighted.
(B) miR-7 targeted genes were sorted according to the number of chimeric reads supporting them. Top targets for miR-7 are non-coding RNAs: CDR1-AS and Cyrano. The miRNA:gene pairs involving a non-coding target transcript are highlighted.
(C) CDR1-AS circular RNA is densely covered by miR-7 targets. However, the highest coverage peak is for miR-671, which acts as siRNA at this locus. Read coverage is shown on a log-scale. The arrow indicates the head-to-tail circular splice junction, and 5'-3' direction is clockwise. Left panel corresponds to the human brain, right is for the mouse brain.
(D) The only miR-7 binding site on Cyrano resides in the middle of conserved region supporting functional importance of miR-7:Cyrano interaction. PhyloP score was calculated for UCSC 100 vertebrates multi-species alignment.
(E) miR-7 shows almost perfect complementarity to its binding site on Cyrano according to in silico hybridization done by RNAhybrid. As 10th and 11th nucleotide of the miRNA are not paired, miRISC cannot cut Cyrano at this spot. Conversely, the targeted transcript may destroy miR-7 loaded to Ago.

While exploring the interactions between the miRNAs and transcripts expressed in human brain we came across a strong association between miRNAs from let-7 family and genes residing on mitochondrial DNA (mtDNA). Despite relatively small number of genes encoded by mtDNA, it occurred as the most targeted chromosome (we denote mtDNA as chromosome here for consistency, even though it lacks chromatin) by let-7 family (Fig. 20A). This might be caused by the over-representation of mitochondrial transcripts in the recovered interactions. In contrast to this assumption, we observed a specific enrichment for let-7 among miRNAs bound to mitochondrial transcripts (Fig. 20B). Moreover, mtDNA was quite uniformly covered with the binding sites for eight different members of let-7 family (Fig. 20C), which excludes that this phenomenon was merely a sequencing or/and mapping artifact. For the AGO-CLIP experiments performed in human heart (Spengler et al., 2016) and mouse brain (Moore et al., 2015) we also detected a strong interplay between mitochondrial RNAs and let-7 family (Fig. 20D,E). As for human brain, these experiments were executed in post-mortem tissue. In contrast, we did not observe any enrichment of let-7 binding sites on mtDNA in the previously analyzed datasets (table 1) performed in cell cultures. Thus, it is tempting to attribute the let-7:mitochondria association to hypoxia or apoptosis processes, which accompany the preparation of post-mortem samples.

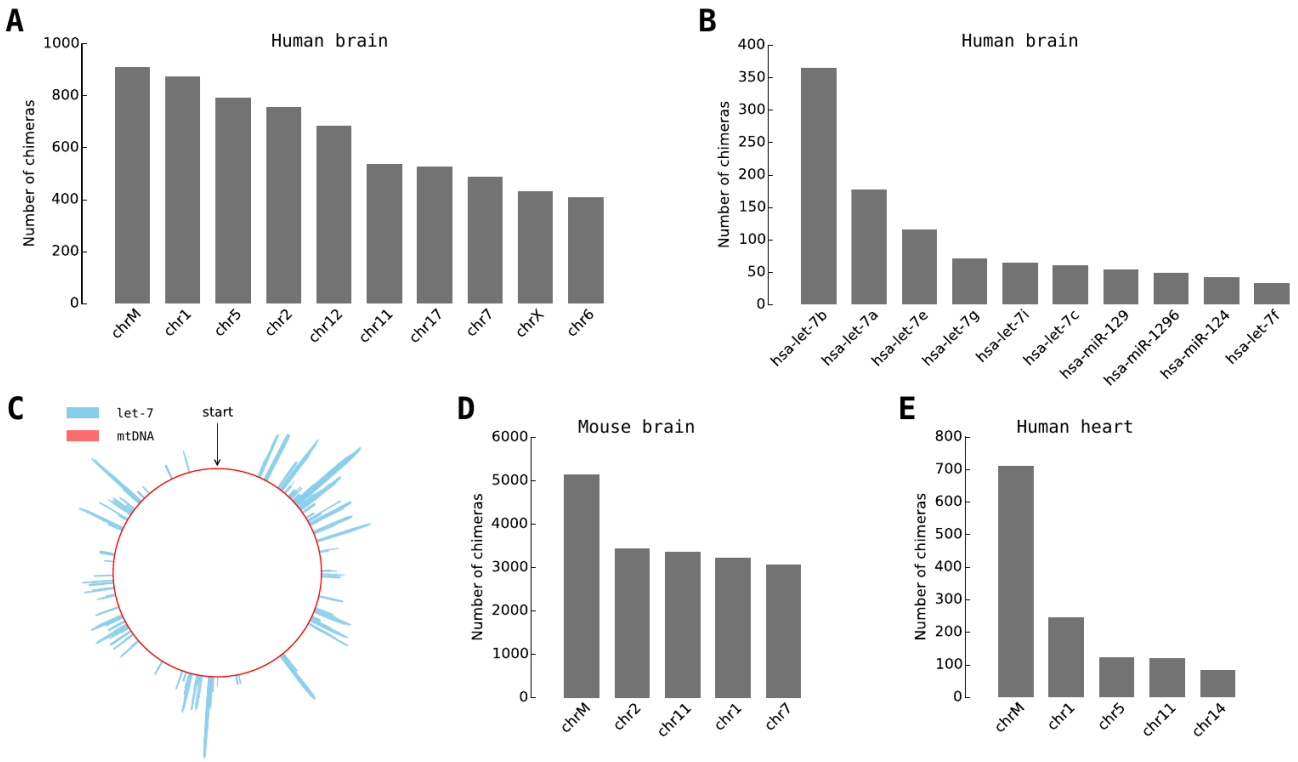

Figure 20| MiR-7, miR-124, CDR1-AS and Cyrano may constitute a regulatory interplay in mammalian brain

(A) Targets of let-7 miRNA family members were grouped according to the host chromosome. Mitochondrial 'chromosome' (chrM) appeared the most targeted, even though it is greatly shorter than others. The analysis was performed for miRNA:targets in human brain.
(B) Eight out of ten top abundant miRNAs in miRNA:mtRNA interactions involve a member of let-7 family. Analysis was performed for miRNA:targets in human brain.
(C) miRNAs from let-7 family have multiple binding sites on mitochondrial transcripts. Read coverage is shown on a log-scale. The arrows indicate the start position of mitochondrial DNA, and 5'-3' direction is clockwise. Analysis was performed for miRNA:targets in human brain.
(D) The same analysis as in (A) was done for miRNA:targets in mouse brain.
(E) The same analysis as in (A) was performed for miRNA:targets in human heart (Spengler et al., 2016)

5. Discussion

5.1 PAR-CLIP experiments augmented with a ligation step are able to retrieve true endogenous miRNA:target interactions with a high specificity

A miRNA acts as a gene suppressor and its functionality depends on the composition of the set of available targets. In other words, a gene expression pattern of a cell in a particular state (cell type, stress, cell cycle and so on) determines biological role of a miRNA. Moreover, miRNAs may depend on the expression of themselves. Indeed, in many cases one gene can be bound by more than one individual miRNA and this combinatorial targeting is synergistic for the repression. Altogether it means that the exploration of miRNAs in a variety of biological processes benefits from being context-specific. AGO-CLIP experiments address this point, uncovering miRNA binding sites in the given cellular contexts. However, additional computational efforts are required to assign a unique miRNA to each binding site. These predictions typically rely on a presence of a seed match and a number of binding sites remain orphans. It is also impossible to distinguish the targets of miRNAs from the same family, since they share the seed sequence. In order to overcome these limitations we developed a strategy to generate and identify reads harboring both miRNA and its cognate target, so-called chimeras. Here I admire the work performed in David Tollervey lab (Kudla et al., 2011), which inspired our research. We introduced an additional ligation step into AGO-CLIP protocol and generated a considerable number of miRNA:target chimeras. Moreover, our experiment also provided a deep compendium of Argonaute binding sites in C.elegans (29.000 individual binding site). For these sites we observed an enrichment of miRNA seed matches and high numbers of T:C conversions, which both are hallmarks of decent AGO-CLIP performance. Thus, we generated broader (29,000 vs 4.800 binding sites) and more precise (mean binding site length 42 vs 122) miRNA interaction map in C. elegans compare to the previous approach (Zisoulis et al., 2010).

For the discovered miRNA:target interactions the most important question was whether they indeed represent true endogenous binding events. Therefore, we considered a number of routes which might lead to the erroneous findings. (1) miRNA:targets might be conventional reads which were incorrectly mapped as chimeras. In order to address this problem I developed the filtering strategy (see 4.1 and 4.2) to keep false discovery rate beyond a reasonable threshold (typically 0.05). (2) In a course of experiment a miRNA might be non-specifically ligated to a piece of RNA which is not bound by AGO, leading to the generation of non-endogenous interactions. In contrast to this assumption, we found around 90% of the sequences ligated with miRNAs to reside on AGO binding sites. Moreover, we observed a clear enrichment in basepairing via a seed site and hybridization patterns typical for miRNA targeting. Thus, our findings are very likely originated from AGO:miRNA:target complexes. (3) Are these complexes endogenous? There might be a scenario, where the endogenous targets are disconnected from AGO and replaced by the other sequences in a course of experiment. Since these

non-endogenous targets are also bound according to the miRNA binding rules, they might posses the expected hybridization pattern and might be eagerly mapped to the AGO binding sites. In order to exclude this scenario, we checked the number of T:C conversions, since they are byproducts of covalent bonds between the targets and Argonaute proteins introduced prior the lysis. As expected for the endogenous interactions. The T:C rate occurred to be high in the sequences ligated to miRNAs (but not for the miRNAs' sequences), even higher than for the AGO binding sites. Moreover, we found only a marginal fraction of the interactions involving E.coli RNAs, which are quite abundant in the lysis, since E.coli is a food source for C.elegans. Thus, we can conclude that the substantial fraction of the discovered miRNA:targets are indeed true endogenous interactions.

5.2 miRNA:target chimeras can be generated in conventional AGO-CLIP experiment

To our surprise we detected a comparable amount of chimeric reads in the control samples, where no ligase was added in course of experiment. Furthermore, these unexpected miRNA:target chimeras largely overlapped with those from 'ligation' samples and possessed all the hallmarks of true endogenous binding events. Generally speaking, it was impossible to distinguish the interactions from 'control' or 'ligation' experiment. However, we found one very specific difference. The chimeras coming from 'control' samples harbored cut miRNA part while there was a mixture of cut and complete miRNAs inside the chimeras from 'ligation' samples. The fraction of the chimeras with shortened miRNAs was almost equal for both experiments. Thus, we assumed that there are two different agents which are able to produce miRNA:target chimeras. First is T4 RNA ligase which generates chimeric reads with a complete miRNA part in agreement with substrate specificity of T4 ligase, as it theoretically cannot use the cut 3'end of a miRNA. The second is a mysterious endogenous ligase, which is active in the lysis, at least for the CLIP experiment. Being an enzyme, this ligase might be substrate-specific. Indeed, we observed a strong enrichment in guanosine upstream the linked chimeric parts. A tRNA ligase, coined HSPC117 in human, fits well to our suspect description. It was shown to link the ends cut by RNase T1, which agrees with our findings (Popow et al., 2011). Moreover it is well-conserved among eukaryotes and hence most likely has a homologue in *C.elegance*. However, further knock-out or knock-down experiments are required to prove the engagement of HSPC117 homologues in chimera-generation process.

tRNA ligase is widespread among eukaryotes and potentially can generate miRNA:target chimeras in AGO-CLIP experiments performed for a variety of species. Since the reads from CLIP sequencing are typically mapped globally (that is the whole sequence of a read must be aligned), the chimeras are discarded as failed to be mapped. They might be also interpreted as single miRNAs or AGO targets if the soft-clipping was utilized for the mapping. In contrast, our pipeline was able to discover around

14.000 thousands interactions in nine already published AGO-CLIP datasets. Further, we analyzed each AGO-CLIP sequencing dataset as it appeared in public repositories. In total, we discovered around 40.000 interactions for human, mouse, C.elegans and human cells infected with herpesviruses. These interactions were also paired according to the known miRNA binding rules and harbored elevated number of characteristic nucleotide conversions. Remarkable, both HITS-CLIP and PAR-CLIP methodologies were able to generate comparable amounts of chimeric reads. Recently thousands of miRNA:target chimeras were also discovered via iCLIP method (individual-nucleotide resolution cross-linking immunoprecipitation) (Broughton et al., 2016).

For some of the discovered interactions we had a chance to test their functional relevance, as miRNA perturbations were available for the corresponding biological systems. We found that genes harboring interactions via seed site in general respond stronger to a perturbation of a cognate miRNA than those which have a seed match but were not supported by the chimeras. That is, binding sites involved in chimeric reads occurred to be more functionally potent than those found via pure sequence prediction. Furthermore, the interactions via imperfect seed pairing also occurred to have a repressive potential. Because of low information content these imperfect miRNA recognition motifs cannot be predicted by computational approaches. Thus, the analyses of chimeras revealed a broad layer of functionally relevant but yet hidden miRNA binding sites.

It is also possible to test functional importance of cis-regulatory elements based on their relative conservation. The seed match sequences engaged in the miRNA:target interactions occurred to be more conserved than those predicted in 3'UTRs or AGO binding sites. There can be two possible explanations of this result. (1) As a probability for a particular 6 or 7 mer to occur by chance is not negligible (4**6 or 4**7), some seed matches in the 3'UTRs or AGO binding sites might be predicted erroneously. (2) As conserved seed matches are generally more functionally important than non-conserved, their context can be more suitable to allow a stable, long-lasting binding of a miRNA. Consequently, these interactions have higher chances to emerge as chimeric reads. The same logic can be applied for the AGO binding sites, so why do we see a difference? In a typical AGO CLIP experiment millions of reads can be mapped continuously with only few thousands being chimeras. Therefore weak and transient interactions might be supported by a number of reads enough to form a valid binding site. In contrast, as only a small fraction of target sites occur in chimeras, the selection based on their affinity becomes more specific. Thus, chimeras might be enriched in more stable and hence more conserved binding events. This hypothesis is also supported by the elevated number of miRNA:target interactions via 3'UTRs comparing to AGO binding sites, which are supposed to be more durable than those in CDS (Fang and Rajewsky, 2011; Gu et al., 2009).

Imperfect seed matches also occurred to be specifically selected in course of evolution, which supports their functional significance. Strikingly, the mismatches inside the imperfect seeds were also

conserved. That is, imperfect recognition motifs are not byproducts of the degradation or construction of the perfect ones. Instead, for a great number of genes it was evolutionary beneficial to have a miRNA binding site but with decreased repressive potency. According to biological reductionism, it is better to have one strong cis-regulatory site, than two weak with the same functional outcome. However, multiple imperfect recognition motifs might be beneficial for combinatorial miRNA binding to ensure a contextual fine-tuning of gene expression.

5.3 Chimeric reads allow deeper understanding of miRNA binding rules

Bioinformatics tools typically require a seed match or its variations to detect miRNA binding sites. Therefore, their predictions constrained to the already predefined binding modes and cannot be used to explore miRNA binding in an unbiased way. Indeed, exploring a contribution of a seed region to the miRNA binding sites, which were selected to have it, is like opening a safe with a key which lies inside. In contrast, miRNA:target chimeras do not rely on any assumption on recognition patterns. If we were completely ignorant to all the previous knowledge on miRNA binding, we would be able to determine the importance of the seed region solely based on the discovered chimeras. Moreover, the direct assignment of a miRNA to its binding site allowed us to explore the contribution and positioning of the mismatches inside the seed.

We found that the majority (~80%) of miRNA:targets are paired via seed (nucleotides 2-8 with one mismatch allowed), rendering this region as a key recognition motif. We analyzed the positional frequencies of the mismatches to the miRNA seed, and found nucleotides 3-6 to particularly avoid them. This was not a surprise, since mismatches in the central nucleotides disrupt the stacking interactions in miRNA:target duplex. The nucleotides 2 and 7 are also, but less, eagerly paired, while the contribution of the 8th miRNA base seems to be less important. Indeed, only 60% of the interactions having a 2-7 seed match also involve the 8th nucleotide, pointing the auxiliary role of the latter. However, miRNA binding rules seem to be quite flexible. In a number of cases the pairing via the 8th nucleotide may compensate the unpaired 2nd, creating a previously described “offset” seed match (nucleotides 3-8). It is important to remark that here we discuss the general trends in miRNA binding, which may not be applicable for an individual miRNA. For example, we discovered only a subset of miRNAs which bind their targets via a 3’end, which is not the case for all miRNAs. Even more extreme example is viral miR-K3 which preferentially utilizes an offset seed instead of a normal one.

Our findings regarding miRNA binding generally agree with AGO structural studies. According to the work performed by Schirle and colleagues (Schirle et al., 2014) a miRNA downloaded into Argonaute protein is arranged in a way to facilitate the binding to the nucleotides 2-5. The binding via

this region initiates conformational changes that expose nucleotides 6 and 7 for further pairing. Nucleotides at the positions 8 and 13-16 also become accessible but less eagerly paired, while the central region of a miRNA is barely reachable by a target RNA. This scheme explains high fraction of the discovered miRNA:target interactions via 2-7 seed with an auxiliary pairing to the 8th nucleotide. For the miRNAs extensively targeting via 3'end binding, 13-16 nt region is almost always paired with a target.

We found that a majority (~60%) of miRNA:target interactions via a seed match have an adenosine in front of the 1st nucleotide of a miRNA. This evidence is in line with 3-fold increased affinity for adenosines opposite miRNA 5'end to the conserved binding pocket in Argonaute protein (compare to the other nucleotides). We haven't observe a particular preferences in nucleotide composition of the mismatches inside the seeds, including G:U pairs. Guanosine and uridine were shown to hybridize with an energy similar to Watson-Crick base pairs (Mizuno and Sundaralingam, 1978). However the geometry of this pair distorts the stacking interactions inside the seed:seed-match duplex rendering its unfavorable for miRNA targeting (Schirle et al., 2014). In line with this evidence G:U pairs are not commonly found among the predicted Ago binding sites (Brennecke et al., 2005) and in the discovered miRNA:target interactions.

There were two major disagreements with the structural study performed by Schirle and colleagues. (1) We concluded that the core seed is a stretch of nucleotides 3-6, while the authors associate initial recognition to nucleotides 2-5. (2) We found nucleotide 9 to be bound quite often, while it has a low accessibility in the predicted AGO:miRNA:RNA complex. However, the structural peculiarities of Ago may not be directly translated into the captured interactions. As binding rules derived from miRNA:target chimeras are a direct product of the interaction kinetics but not the structural constrains, we cannot demand their one-to-one correspondence to those derived from structural studies. It might be that the initiation via 3-5 region is enough to cause nucleotide 6th exposure and the affinity of a target RNA to 2-5 region is generally lower than to nucleotides 3-6.

One fifth of the discovered miRNA:target interactions lack detectable pairing via seed-match. As seedless interactions also harbor strong crosslinking signatures and tend to reside on 3'UTRs of protein coding genes, they are unlikely enriched in false positives. These interactions rather use week seed-pairing with a couple of bulges and/or rely on extensive 3' end binding. These targeting modes were also described previously (Helwak et al., 2013). It also has to be remarked that a significant number of miRNAs undergoes post-transcriptional editing (Telonis et al., 2015), which may change their seed sequence comparing to the genomic one. Consequently, the seed-pairing for the edited miNRAs might be underestimated. Perhaps, some of the seedless interactions may arise from the Ago binding independent of miRNA sequence. For example, Smaug protein was shown to bind mRNAs and transfer them to Ago for further repression (Pinder and Smibert, 2013).

5.4 miRNAs in mammalian brain

Post-transcriptional gene regulation is known to be amazingly complex and important in neuronal tissue. Therefore, miRNA:target interactions discovered in human (~ 23 000 interactions) and mouse brain (~ 180 000 interactions) were of particular interest for us. Similar to the miRNA:targets in cell cultures and C.elegans, these interactions mainly involve protein coding genes and heavily rely on pairing via seed region. However we observed the difference in the binding sites' distribution along the mRNA with a brain-specific emphasis on targeting coding sequence rather than 3'UTR. As circular RNAs are abundant in human brain, a number of miRNA:CDS interactions can be indeed miRNA:circRNA interactions. Despite the observed enrichment of CDS binding sites among the circles expressed in brain, the fraction of targets on coding sequence of mRNAs, which does not have a detected circular isoform, remains unexpectedly high. Moreover, circles themselves tend to be composed of CDS exons, as the last 3'UTR exon cannot be spliced. Thus, one may consider an enhanced miRNA targeting via coding sequence to be brain-specific.

Another peculiarity of the analyzed datasets was a strong association of miRNAs from let-7 family with mitochondrial transcripts. Being only a tiny fraction of the expressed genes, mitochondrial mRNAs attracted more let-7 miRNAs than transcripts from any other chromosome (we denote mtDNA as chromosome here for consistency, even though it lacks chromatin). Furthermore, mtDNA occurred to be densely covered by eight different members of let-7 family but not any other miRNAs. Therefore, it is unlikely that this phenomenon was caused by mapping artifacts. Unfortunately, there is still a lack of evidences of miRNA actions inside mitochondria, so further experimental validation of our findings is in demand. The profound let-7:mtRNA association was also found for the Ago-CLIP experiment performed in human heart (~ 3 000 interactions), but not for the experiments done in *C.elegans* and cell cultures. So far we have found only one peculiarity which sets apart the datasets with multiple let-7:mtRNA interactions from the others: these datasets are originated from the experiments performed in post-mortem tissue. Therefore, we can speculate that this phenomenon may be linked to the hypoxia and/or apoptosis, which accompany the preparation of post-mortem samples.

As chimeras provide us an information about genuine miRNA:target pairs, we were able to resolve an intriguing interplay between non-coding RNAs in mammalian brain. MiR-7 is known to be important in brain function and development. Recently it was shown to be sponged by circular RNA CDR1-AS which harbors dozens of seed matches for miR-7 (~70 in human and ~ 100 in mice). In line with this evidence we found that 90% of miR-7 interactions involve CDR1-AS, rendering the latter as almost exclusive target. However, on a level of individual binding sites miR-7 was not the top interactor of CDR1-AS. Indeed, miR-671 binds to a locus on this circRNA with almost full complementarity, and

this interaction is supported by a significant number of the chimeric reads. Being complementary to its binding site miR-671 can cut CDR1-AS, which was previously observed by Hansen and colleagues (Hansen et al., 2011). The slicing potential of miR-671 is of particular importance for CDR1-AS turnover, as circular RNAs cannot be degraded by endonucleases. Another brain-specific interaction with almost full complementarity involves aforementioned miR-7 and extremely well-conserved locus on non-coding RNA Cyrano. In contrast to miR-671:CDR1-AS the 9th and 10th nucleotides of the miRNA are not paired, which obstructs its slicing activity. Instead, similar binding architecture was recently reported to cause the degradation of a miRNA but not the targeted transcript (de la Mata et al., 2015). Remarkably, Cyrano is the second top target for miR-7 after CDR1-AS, and CDR1-AS is the top target for miR-671. Thus, the top-supported by chimeras binding sites for the brain-specific miR-7 and miR-671 are circRNA CDR1-AS and non-coding RNA Cyrano.

What can be the functional consequences of the aforementioned interplay? MiR-7 was shown to be downregulated upon CDR1-AS knock-down, while the expression of miR-671 increases (Piwecka et al., 2017). Therefore, I can speculate that CDR1-AS protects miR-7 from the Cyrano, which destroys it otherwise. In its turn miR-671 may slice CDR1-AS which leads to the release of multiple miR-7:Ago complexes. Therefore spatial and temporal expression of miR-7 depends on CDR1-AS, Cyrano, and miR-671, which may allow precise and robust control over miR-7 targets. Remarkably, this regulatory circuit is present in neuronal tissue, the tissue known for its complex post-transcriptional gene regulation. Recently it was shown that single-synapse stimulation may raise the rate of local miRNA maturation (for miR-181a) (Sambandan et al., 2017). If the local expression of miR-671 can be also triggered by a synapse stimulation, it may result in the enhanced slicing of CDR1-AS and subsequent release of dozens miR-7:Ago complexes. Thus, the initial signal may be significantly amplified via miR-671:CDR1-AS:miR-7 loop leading to a local burst in miR-7 expression. So far I can only speculate on these potential regulatory mechanisms, but I believe that they deserve a thorough experimental investigation.

5.5 Potential applications and limitations of chimera-based methods

The main advantage of chimera-based approaches is their ability to provide context-dependent, direct miRNA target interactions for a given biological system. This information can be valuable for systematic exploration of post-transcriptional gene regulation, and also for more specific research on a particular miRNA function in disease and development. Indeed, many miRNAs were observed to be misexpressed in a variety of pathological processes. The exploration of their direct targets may uncover the regulatory pathways important for the potential medical applications. Here we analyzed medically-relevant systems: human cell lines infected with Epstein-Barr and Sarcoma Kaposi

herpesviruses. The problem of the identification of miRNAs' binding sites was even complicated in this case, as herpesviruses encode their own miRNAs. Consequently, the binding sites for the viral miRNAs do not have to be conserved, which obstructs their identification. Moreover, some of the viral miRNAs share the seed sequence with the host ones, hence their predicted seed-matches are indistinguishable (e.g. KSHV miR-K11 mimics human miR-155). The direct identification of miRNA:target interactions in virus-infected cells allowed to overcome these specific obstacles, and generated a comprehensive map of viral miRNAs' bindings sites on the host transcripts. The information of viral genomes is rather condensed (that is, most of the sequence is exonic) most probably due to evolutionary pressure on a size of viral particles. Indeed, the smaller the virus, the easier it can invade a host cell. Consequently, each transcript encoded in viral genome, including miRNAs, should be relevant for a virus infection and/or maintenance. Therefore, the analysis of the human transcripts targeted by viral miRNAs may uncover pathogen strategies important for survival and proliferation.

Gigh specificity of the discovered miRNA-target also allows revealing regulatory circuits involved in normal function and development. Apart from the extensively discussed miR-7 interplay with other non-coding RNAs in mammalian brain, we also observed previously described negative feedback loop of miRNA biogenesis pathway (Forman et al., 2008). Indeed, Argonaute transcripts participate in the interactions with multiple miRNAs (Fig. d1). As the expression levels of miRNAs positively correlate with the number of miRISC complexes, Argonaute is negatively-regulated by itself. Thus, this negative feedback loop may compensate temporal spikes in the translational rates of Ago and miRNAs targeting it.

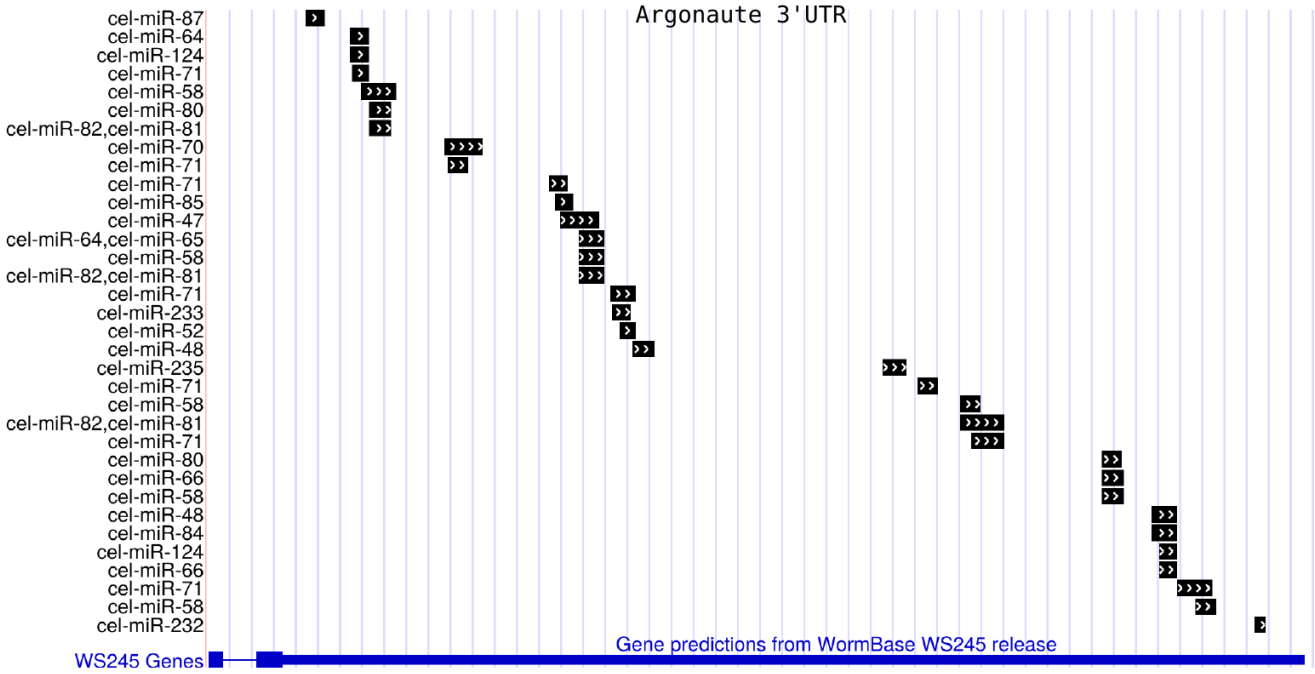

Figure d1| Argonaute transcripts are regulated by miRNAs via negative feedback

(A) Argonaute 3'UTR in *C. elegans* harbors multiple binding sites for a variety of miRNAs. Alg-1 3'UTR is annotated according to the WormBase version WS245 (Howe et al., 2016).

In order to quantitatively assess functional outcomes of miRNA regulatory circuits, the number of chimeric reads supporting miR-A:gene-B interaction must reflect the expression of a miR-A and gene-B with a decent precision. As miRNA:target chimeras represent only a small fraction of the total sequencing pool, and the ligation efficiency may be skewed by multiple biases, we did not anticipate a strong correlation between miRNAs expression levels and their occurrences in chimeras. Indeed, for the datasets with only few thousands chimeric reads (cell lines datasets) the correlation was relatively weak (Pearson correlation coefficient was constantly lower than 0.4). However, for the dataset with around 100.000 miRNA:target chimeras (human brain dataset) the correlation became reasonably high (Pearson correlation coefficient 0.74). Unfortunately, we cannot reliably assess how well mRNAs expression levels correlate with their occurrences in chimeras, as the latter numbers are in general low even for highly-expressed transcripts. Therefore, further improvements in chimera generation efficiency are required to build reliable quantitative models of miRNA regulation.

An efficient generation of miRNA:target chimeras depends on the following steps. (1) The target sequence has to be covalently crosslinked to an Argonaute protein. As we observed the crosslink predominantly right upstream the seed match, only a few nucleotides in the target sequence can be efficiently crosslinked to the protein. This preference for the position of crosslink lowers the efficiency of chimera generation in general and particularly for those with target nucleotides upstream the seed-match paired with miRNA sequence, as these nucleotides are constrained in forming other chemical bonds. The aforementioned problem becomes even more exaggerated for PAR-CLIP protocols, since crosslink can be generated only for uridines in target sequence. (2) An RNase has to cut the target and the miRNA sequence in a way suitable for the ligation reaction. That is, the cleaved RNA ends have to be appropriate substrates for the ligation agent (so far not fully characterized). Moreover the 3' end of a miRNA and 5' end of a cognate target have to be in close proximity to each other. (3) The lengths of both miRNA and target part need to be long enough to be further identified as chimeras. (4) All the multiple biases and complications related to the AGO-CLIP experiments also influence the generation of chimeras. Thus, the chimera-generation process is complex and prone to the biases. Indeed, for a number of AGO-CLIP datasets we failed to find even a single reliable miRNA:target chimera.

Despite the aforementioned concerns we found around a hundred of thousands miRNA:target chimeras in human brain, and the numbers of chimeric reads correlated well with miRNA expression levels. That is the generation of these chimeras was not significantly biased among miRNAs. Furthermore, the additional improvements in the AGO-CLIP protocol already have led to the generation of tens of thousands miRNA:target chimeras in mouse brain (Moore et al., 2015) and *C. elegans* (Broughton et al., 2016). Therefore, further boosts in the sensitivity of miRNA:target

interactions recovery can be expected in the next few years. As for the specificity, Broughton and colleagues proposed rather simple PCR-based strategy to confirm the selected interactions (Broughton et al., 2016).

AGO-CLIP experiments are relatively complex, expensive, and require an extensive tissue preparation. Moreover, for many biological systems it is cumbersome to establish an uptake of modified nucleotides and generate an antibody for Argonaute protein. Therefore, chimera-based approach cannot fully substitute miRNA prediction tools. In fact, the latter can benefit from the knowledge on miRNA binding derived from the analysis of miRNA:target interactions. Thus, a chimera-based approach is valuable for the context-specific discovery of miRNA:targets, and also for the expansion of our knowledge on miRNA biology in general.

Bibliography

An, X.P., Hou, J.X., Li, G., Song, Y.X., Wang, J.G., Chen, Q.J., Cui, Y.H., Wang, Y.F., and Cao, B.Y. (2012). Polymorphism identification in the goat KITLG gene and association analysis with litter size. Anim Genet *43*, 104–107.

Aravind, L., Watanabe, H., Lipman, D.J., and Koonin, E.V. (2000). Lineage-specific loss and divergence of functionally linked genes in eukaryotes. Proc Natl Acad Sci U S A *97*, 11319–11324.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25–29.

Aukerman, M.J., and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. Plant Cell *15*, 2730–2741.

Aviv, T., Lin, Z., Ben-Ari, G., Smibert, C.A., and Sicheri, F. (2006). Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1p. Nat Struct Mol Biol *13*, 168–176.

Barbosa, C., Peixeiro, I., and Romão, L. (2013). Gene expression regulation by upstream open reading frames and human disease. PLoS Genet *9*, e1003529.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215–233.

Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C., et al. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J *33*, 981–993.

Berkovits, B.D., and Mayr, C. (2015). Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. Nature *522*, 363–367.

Boss, I.W., Nadeau, P.E., Abbott, J.R., Yang, Y., Mergia, A., and Renne, R. (2011). A Kaposi's sarcoma-associated herpesvirus-encoded ortholog of microRNA miR-155 induces human splenic B-cell expansion in NOD/LtSz-scid IL2Rγnull mice. J Virol *85*, 9877–9886.

Botto, S., Totonchy, J.E., Gustin, J.K., and Moses, A.V. (2015). Kaposi Sarcoma Herpesvirus Induces HO-1 during De Novo Infection of Endothelial Cells via Viral miRNA-Dependent and -Independent Mechanisms. MBio *6*, e00668.

Boudreau, R.L., Jiang, P., Gilmore, B.L., Spengler, R.M., Tirabassi, R., Nelson, J.A., Ross, C.A., Xing, Y., and Davidson, B.L. (2014). Transcriptome-wide discovery of microRNA binding sites in human brain. Neuron *81*, 294–305.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. (2005). Principles of microRNA-target recognition. PLoS Biol *3*, e85.

Broughton, J.P., and Pasquinelli, A.E. (2013). Identifying Argonaute binding sites in Caenorhabditis elegans using iCLIP. Methods *63*, 119–125.

Broughton, J.P., Lovci, M.T., Huang, J.L., Yeo, G.W., and Pasquinelli, A.E. (2016). Pairing beyond the Seed Supports MicroRNA Targeting Specificity. Mol Cell *64*, 320–333.

Brümmer, A., and Hausser, J. (2014). MicroRNA binding sites in the coding region of mRNAs: extending the repertoire of post-transcriptional gene regulation. Bioessays *36*, 617–626.

Cai, X., Hagedorn, C.H., and Cullen, B.R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA *10*, 1957–1966.

Calin, G.A., Dumitru, C.D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., et al. (2002). Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A *99*, 15524–15529.

Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A *106*, 7507–7512.

Catalanotto, C., Cogoni, C., and Zardo, G. (2016). Microrna in control of gene expression: an overview of nuclear functions. Int J Mol Sci *17*.

Cerutti, H., and Casas-Mollano, J.A. (2006). On the origin and functions of RNA-mediated silencing: from protists to man. Curr Genet *50*, 81–99.

De Chevigny, A., Coré, N., Follert, P., Gaudin, M., Barbry, P., Béclin, C., and Cremer, H. (2012). miR-7a regulation of Pax6 controls spatial origin of forebrain dopaminergic neurons. Nat Neurosci *15*, 1120–1126.

Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature *460*, 479–486.

Chi, S.W., Hannon, G.J., and Darnell, R.B. (2012). An alternative mode of microRNA target recognition. Nat Struct Mol Biol *19*, 321–327.

Chou, C.-H., Chang, N.-W., Shrestha, S., Hsu, S.-D., Lin, Y.-L., Lee, W.-H., Yang, C.-D., Hong, H.-C., Wei, T.-Y., Tu, S.-J., et al. (2016). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res *44*, D239–47.

Cimmino, A., Calin, G.A., Fabbri, M., Iorio, M.V., Ferracin, M., Shimizu, M., Wojcik, S.E., Aqeilan, R.I., Zupo, S., Dono, M., et al. (2005). miR-15 and miR-16 induce apoptosis by targeting BCL2. Proc Natl Acad Sci U S A *102*, 13944–13949.

Corcoran, D.L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R.L., Keene, J.D., and Ohler, U. (2011). PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. Genome Biol *12*, R79.

Dahlke, C., Maul, K., Christalla, T., Walz, N., Schult, P., Stocking, C., and Grundhoff, A. (2012). A microRNA encoded by Kaposi sarcoma-associated herpesvirus promotes B-cell expansion in vivo. PLoS ONE *7*, e49435.

Denli, A.M., Tops, B.B.J., Plasterk, R.H.A., Ketting, R.F., and Hannon, G.J. (2004). Processing of primary microRNAs by the Microprocessor complex. Nature *432*, 231–235.

Derti, A., Garrett-Engele, P., Macisaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M., and Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. Genome Res *22*, 1173–1183.

Dodt, M., Roehr, J.T., Ahmed, R., and Dieterich, C. (2012). FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. Biology *1*, 895–905.

El Hajj, P., Gilot, D., Migault, M., Theunis, A., van Kempen, L.C., Salés, F., Fayyad-Kazan, H., Badran, B., Larsimont, D., Awada, A., et al. (2015). SNPs at miR-155 binding sites of TYRP1 explain discrepancy between mRNA and protein and refine TYRP1 prognostic value in melanoma. Br J Cancer *113*, 91–98.

Fang, Z., and Rajewsky, N. (2011). The impact of miRNA target sites in coding sequences and in 3'UTRs. PLoS ONE *6*, e18067.

Forman, J.J., Legesse-Miller, A., and Coller, H.A. (2008). A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. Proc Natl Acad Sci U S A *105*, 14879–14884.

Fred, R.G., Mehrabi, S., Adams, C.M., and Welsh, N. (2016). PTB and TIAR binding to insulin mRNA 3'- and 5'UTRs; implications for insulin biosynthesis and messenger stability. Heliyon *2*, e00159.

Friedländer, M.R., Mackowiak, S.D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res *40*, 37–52.

Friedman, R.C., Farh, K.K.-H., Burge, C.B., and Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res *19*, 92–105.

Garcia, D.M., Baek, D., Shin, C., Bell, G.W., Grimson, A., and Bartel, D.P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol *18*, 1139–1146.

Gorgoni, B., and Gray, N.K. (2004). The roles of cytoplasmic poly(A)-binding proteins in regulating gene expression: a developmental perspective. Brief Funct Genomic Proteomic *3*, 125–141.

Gottwein, E. (2012). Kaposi's Sarcoma-Associated Herpesvirus microRNAs. Front Microbiol *3*, 165.

Gottwein, E., Mukherjee, N., Sachse, C., Frenzel, C., Majoros, W.H., Chi, J.-T.A., Braich, R., Manoharan, M., Soutschek, J., Ohler, U., et al. (2007). A viral microRNA functions as an orthologue of cellular miR-155. Nature *450*, 1096–1099.

Gottwein, E., Corcoran, D.L., Mukherjee, N., Skalsky, R.L., Hafner, M., Nusbaum, J.D., Shamulailatpam, P., Love, C.L., Dave, S.S., Tuschl, T., et al. (2011). Viral microRNA targetome of KSHV-infected primary effusion lymphoma cell lines. Cell Host Microbe *10*, 515–526.

Grundhoff, A., Sullivan, C.S., and Ganem, D. (2006). A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses. RNA *12*, 733–750.

Gu, S., Jin, L., Zhang, F., Sarnow, P., and Kay, M.A. (2009). Biological basis for restriction of microRNA targets to the 3' untranslated region in mammalian mRNAs. Nat Struct Mol Biol *16*, 144–150.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., et al. (2010). PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins. J Vis Exp.

Halees, A.S., El-Badrawi, R., and Khabar, K.S.A. (2008). ARED Organism: expansion of ARED reveals AU-rich element cluster variations between human and mouse. Nucleic Acids Res *36*, D137–40.

Hammond, S.M., Bernstein, E., Beach, D., and Hannon, G.J. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. Nature *404*, 293–296.

Hansen, T.B., Wiklund, E.D., Bramsen, J.B., Villadsen, S.B., Statham, A.L., Clark, S.J., and Kjems, J. (2011). miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. EMBO J *30*, 4414–4422.

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell *153*, 654–665.

Hitti, E., Bakheet, T., Al-Souhibani, N., Moghrabi, W., Al-Yahya, S., Al-Ghamdi, M., Al-Saif, M., Shoukri, M.M., Lánczky, A., Grépin, R., et al. (2016). Systematic Analysis of AU-Rich Element Expression in Cancer Reveals Common Functional Clusters Regulated by Key RNA-Binding Proteins. Cancer Res *76*, 4068–4080.

Howe, K.L., Bolt, B.J., Cain, S., Chan, J., Chen, W.J., Davis, P., Done, J., Down, T., Gao, S., Grove, C., et al. (2016). WormBase 2016: expanding to enable helminth genomic research. Nucleic Acids Res *44*, D774–80.

Hutvágner, G., McLachlan, J., Pasquinelli, A.E., Bálint, E., Tuschl, T., and Zamore, P.D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science *293*, 834–838.

Jens, M. (2016). A Pipeline for PAR-CLIP Data Analysis. Methods Mol Biol *1358*, 197–207.

Jin, H.Y., Oda, H., Lai, M., Skalsky, R.L., Bethel, K., Shepherd, J., Kang, S.G., Liu, W.-H., Sabouri-Ghomi, M., Cullen, B.R., et al. (2013). MicroRNA-17~92 plays a causative

role in lymphomagenesis by coordinating multiple oncogenic pathways. EMBO J *32*, 2377–2391.

Johnson, S.M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., Labourier, E., Reinert, K.L., Brown, D., and Slack, F.J. (2005). RAS is regulated by the let-7 microRNA family. Cell *120*, 635–647.

Jungkamp, A.-C., Stoeckius, M., Mecenas, D., Grün, D., Mastrobuoni, G., Kempa, S., and Rajewsky, N. (2011). In vivo and transcriptome-wide identification of RNA binding protein target sites. Mol Cell *44*, 828–840.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature *421*, 231–237.

Kanellopoulou, C., Muljo, S.A., Kung, A.L., Ganesan, S., Drapkin, R., Jenuwein, T., Livingston, D.M., and Rajewsky, K. (2005). Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. Genes Dev *19*, 489–501.

Kerpedjiev, P., Frellsen, J., Lindgreen, S., and Krogh, A. (2014). Adaptable probabilistic mapping of short reads using position specific scoring matrices. BMC Bioinformatics *15*, 100.

Khabar, K.S.A. (2010). Post-transcriptional control during chronic inflammation and cancer: a focus on AU-rich elements. Cell Mol Life Sci *67*, 2937–2955.

Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. Cell *115*, 209–216.

Kincaid, R.P., and Sullivan, C.S. (2012). Virus-encoded microRNAs: an overview and a look to the future. PLoS Pathog *8*, e1003018.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat Methods *8*, 559–564.

König, J., Zarnack, K., Luscombe, N.M., and Ule, J. (2012). Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet *13*, 77–83.

Krek, A., Grün, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., et al. (2005). Combinatorial microRNA target predictions. Nat Genet *37*, 495–500.

Kudla, G., Granneman, S., Hahn, D., Beggs, J.D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. Proc Natl Acad Sci U S A *108*, 10010–10015.

Landthaler, M., Yalcin, A., and Tuschl, T. (2004). The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. Curr Biol *14*, 2162–2167.

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S.H., and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. EMBO J *23*, 4051–4060.

Leppek, K., Das, R., and Barna, M. (2017). Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat Rev Mol Cell Biol.

Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15–20.

Linnstaedt, S.D., Gottwein, E., Skalsky, R.L., Luftig, M.A., and Cullen, B.R. (2010). Virally induced cellular microRNA miR-155 plays a key role in B-cell immortalization by Epstein-Barr virus. J Virol *84*, 11670–11678.

Lipchina, I., Elkabetz, Y., Hafner, M., Sheridan, R., Mihailovic, A., Tuschl, T., Sander, C., Studer, L., and Betel, D. (2011). Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response. Genes Dev *25*, 2173–2186.

Liu, G., Zhang, R., Xu, J., Wu, C.-I., and Lu, X. (2015). Functional conservation of both CDS- and 3'-UTR-located microRNA binding sites between species. Mol Biol Evol *32*, 623–628.

Loeb, G.B., Khan, A.A., Canner, D., Hiatt, J.B., Shendure, J., Darnell, R.B., Leslie, C.S., and Rudensky, A.Y. (2012). Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. Mol Cell *48*, 760–770.

Londin, E., Loher, P., Telonis, A.G., Quann, K., Clark, P., Jing, Y., Hatzimichael, E., Kirino, Y., Honda, S., Lally, M., et al. (2015). Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. Proc Natl Acad Sci U S A *112*, E1106–15.

López de Silanes, I., Zhan, M., Lal, A., Yang, X., and Gorospe, M. (2004). Identification of a target RNA motif for RNA-binding protein HuR. Proc Natl Acad Sci U S A *101*, 2987–2992.

Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nat Rev Mol Cell Biol *8*, 479–490.

Ma, J.-B., Yuan, Y.-R., Meister, G., Pei, Y., Tuschl, T., and Patel, D.J. (2005). Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. Nature *434*, 666–670.

Mackereth, C.D., and Sattler, M. (2012). Dynamics in multi-domain protein recognition of RNA. Curr Opin Struct Biol *22*, 287–296.

Manzano, M., Shamulailatpam, P., Raja, A.N., and Gottwein, E. (2013). Kaposi's sarcoma-associated herpesvirus encodes a mimic of cellular miR-23. J Virol *87*, 11821–11830.

De la Mata, M., Gaidatzis, D., Vitanescu, M., Stadler, M.B., Wentzel, C., Scheiffele, P., Filipowicz, W., and Großhans, H. (2015). Potent degradation of neuronal miRNAs induced by highly complementary targets. EMBO Rep *16*, 500–511.

Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol *288*, 911–940.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. Nature *495*, 333–338.

Meyer, S., Temme, C., and Wahle, E. (2004). Messenger RNA turnover in eukaryotes: pathways and enzymes. Crit Rev Biochem Mol Biol *39*, 197–216.

Mi, S., Li, Z., Chen, P., He, C., Cao, D., Elkahloun, A., Lu, J., Pelloso, L.A., Wunderlich, M., Huang, H., et al. (2010). Aberrant overexpression and function of the miR-17-92 cluster in MLL-rearranged acute leukemia. Proc Natl Acad Sci U S A *107*, 3710–3715.

Mizuno, H., and Sundaralingam, M. (1978). Stacking of Crick Wobble pair and Watson-Crick pair: stability rules of G-U pairs at ends of helical stems in tRNAs and the relation to codon-anticodon Wobble interaction. Nucleic Acids Res *5*, 4451–4461.

Moore, M.J., Scheel, T.K.H., Luna, J.M., Park, C.Y., Fak, J.J., Nishiuchi, E., Rice, C.M., and Darnell, R.B. (2015). miRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. Nat Commun *6*, 8864.

Mortensen, R.D., Serra, M., Steitz, J.A., and Vasudevan, S. (2011). Posttranscriptional activation of gene expression in Xenopus laevis oocytes by microRNA-protein complexes (microRNPs). Proc Natl Acad Sci U S A *108*, 8281–8286.

Van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., and Brouns, S.J.J. (2009). CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci *34*, 401–407.

Pak, J., and Fire, A. (2007). Distinct populations of primary and secondary effectors during RNAi in C. elegans. Science *315*, 241–244.

Park, J.-E., Heo, I., Tian, Y., Simanshu, D.K., Chang, H., Jee, D., Patel, D.J., and Kim, V.N. (2011). Dicer recognizes the 5' end of RNA for efficient and accurate processing. Nature *475*, 201–205.

Pesole, G., Liuni, S., Grillo, G., and Saccone, C. (1997). Structural and compositional features of untranslated regions of eukaryotic mRNAs. Gene *205*, 95–102.

Pfeffer, S., Zavolan, M., Grässer, F.A., Chien, M., Russo, J.J., Ju, J., John, B., Enright, A.J., Marks, D., Sander, C., et al. (2004). Identification of virus-encoded microRNAs. Science *304*, 734–736.

Pinder, B.D., and Smibert, C.A. (2013). microRNA-independent recruitment of Argonaute 1 to nanos mRNA through the Smaug RNA-binding protein. EMBO Rep *14*, 80–86.

Place, R.F., Li, L.-C., Pookot, D., Noonan, E.J., and Dahiya, R. (2008). MicroRNA-373 induces expression of genes with complementary promoter sequences. Proc Natl Acad Sci U S A *105*, 1608–1613.

Popow, J., Englert, M., Weitzer, S., Schleiffer, A., Mierzwa, B., Mechtler, K., Trowitzsch, S., Will, C.L., Lührmann, R., Söll, D., et al. (2011). HSPC117 is the essential subunit of a human tRNA splicing ligase complex. Science *331*, 760–764.

Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. RNA *10*, 1507–1517.

Rodriguez, A., Vigorito, E., Clare, S., Warren, M.V., Couttet, P., Soond, D.R., van Dongen, S., Grocock, R.J., Das, P.P., Miska, E.A., et al. (2007). Requirement of bic/microRNA-155 for normal immune function. Science *316*, 608–611.

Ruby, J.G., Jan, C.H., and Bartel, D.P. (2007). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83–86.

Rybak-Wolf, A., Stottmeister, C., Glažar, P., Jens, M., Pino, N., Giusti, S., Hanan, M., Behm, M., Bartok, O., Ashwal-Fluss, R., et al. (2015). Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. Mol Cell *58*, 870–885.

Sambandan, S., Akbalik, G., Kochen, L., Rinne, J., Kahlstatt, J., Glock, C., Tushev, G., Alvarez-Castelao, B., Heckel, A., and Schuman, E.M. (2017). Activity-dependent spatially localized miRNA maturation in neuronal dendrites. Science *355*, 634–637.

Samols, M.A., Hu, J., Skalsky, R.L., and Renne, R. (2005). Cloning and identification of a microRNA cluster within the latency-associated region of Kaposi's sarcoma-associated herpesvirus. J Virol *79*, 9301–9305.

Sandhu, S.K., Fassan, M., Volinia, S., Lovat, F., Balatti, V., Pekarsky, Y., and Croce, C.M. (2013). B-cell malignancies in microRNA Eμ-miR-17~92 transgenic mice. Proc Natl Acad Sci U S A *110*, 18208–18213.

Schirle, N.T., Sheu-Gruttadauria, J., and MacRae, I.J. (2014). Structural basis for microRNA targeting. Science *346*, 608–613.

Schwarz, D.S., Hutvágner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. Cell *115*, 199–208.

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58–63.

Sethupathy, P., Corda, B., and Hatzigeorgiou, A.G. (2006). TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA *12*, 192–197.

Shin, C., Nam, J.-W., Farh, K.K.-H., Chiang, H.R., Shkumatava, A., and Bartel, D.P. (2010). Expanding the microRNA targeting code: functional sites with centered pairing. Mol Cell *38*, 789–802.

Skalsky, R.L., Samols, M.A., Plaisance, K.B., Boss, I.W., Riva, A., Lopez, M.C., Baker, H.V., and Renne, R. (2007). Kaposi's sarcoma-associated herpesvirus encodes an ortholog of miR-155. J Virol *81*, 12836–12845.

Skalsky, R.L., Corcoran, D.L., Gottwein, E., Frank, C.L., Kang, D., Hafner, M., Nusbaum, J.D., Feederle, R., Delecluse, H.-J., Luftig, M.A., et al. (2012). The viral and cellular microRNA targetome in lymphoblastoid cell lines. PLoS Pathog *8*, e1002484.

Sokol, N.S., and Ambros, V. (2005). Mesodermally expressed Drosophila microRNA-1 is regulated by Twist and is required in muscles during larval growth. Genes Dev *19*, 2343–2354.

Soller, M., and White, K. (2005). ELAV multimerizes on conserved AU4-6 motifs important for ewg splicing regulation. Mol Cell Biol *25*, 7580–7591.

Spengler, R.M., Zhang, X., Cheng, C., McLendon, J.M., Skeie, J.M., Johnson, F.L., Davidson, B.L., and Boudreau, R.L. (2016). Elucidation of transcriptome-wide microRNA binding sites in human cardiac tissues by Ago2 HITS-CLIP. Nucleic Acids Res *44*, 7120–7131.

Spriggs, K.A., Stoneley, M., Bushell, M., and Willis, A.E. (2008). Re-programming of translation following cell stress allows IRES-mediated translation to predominate. Biol Cell *100*, 27–38.

Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. (2005). Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. Cell *123*, 1133–1146.

Takamizawa, J., Konishi, H., Yanagisawa, K., Tomida, S., Osada, H., Endoh, H., Harano, T., Yatabe, Y., Nagino, M., Nimura, Y., et al. (2004). Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. Cancer Res *64*, 3753–3756.

Taliaferro, J.M., Lambert, N.J., Sudmant, P.H., Dominguez, D., Merkin, J.J., Alexis, M.S., Bazile, C., and Burge, C.B. (2016). RNA sequence context effects measured in vitro predict in vivo protein binding and regulation. Mol Cell *64*, 294–306.

Teleman, A.A., Maitra, S., and Cohen, S.M. (2006). Drosophila lacking microRNA miR-278 are defective in energy homeostasis. Genes Dev *20*, 417–422.

Telonis, A.G., Loher, P., Jing, Y., Londin, E., and Rigoutsos, I. (2015). Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity. Nucleic Acids Res *43*, 9158–9175.

Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res *33*, 201–212.

Ventura, A., Young, A.G., Winslow, M.M., Lintault, L., Meissner, A., Erkeland, S.J., Newman, J., Bronson, R.T., Crowley, D., Stone, J.R., et al. (2008). Targeted deletion reveals essential and overlapping functions of the miR-17 through 92 family of miRNA clusters. Cell *132*, 875–886.

Yang, W.J., Yang, D.D., Na, S., Sandusky, G.E., Zhang, Q., and Zhao, G. (2005). Dicer is required for embryonic angiogenesis during mouse development. J Biol Chem *280*, 9330–9335.

Yi, R., Qin, Y., Macara, I.G., and Cullen, B.R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev *17*, 3011–3016.

Zhang, L., Huang, J., Yang, N., Greshock, J., Megraw, M.S., Giannakakis, A., Liang, S., Naylor, T.L., Barchetti, A., Ward, M.R., et al. (2006). microRNAs exhibit high frequency genomic alterations in human cancer. Proc Natl Acad Sci U S A *103*, 9136–9141.

Zisoulis, D.G., Lovci, M.T., Wilbert, M.L., Hutt, K.R., Liang, T.Y., Pasquinelli, A.E., and Yeo, G.W. (2010). Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. Nat Struct Mol Biol *17*, 173–179.

Andrews S. (2010) FastQC: a quality control tool for high throughput sequence data. Available: www.bioinformatics.bbsrc.ac.uk/projects/ ? fastqc/Accessed 2012 January. - jourlib.org.

List of publications

Grosswendt, S., Filipchyk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., Gottwein, E., and Rajewsky, N. (2014). Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. Mol Cell 54, 1042–1054.

Piwecka, M., Glažar, P., Hernandez-Miranda, L.R., Memczak, S., Wolf, S.A., Rybak-Wolf, A., Filipchyk, A., Klironomos, F., Cerda Jara, C.A., Fenske, P., et al. (2017). Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. Science 357.

Zappulo, A., van den Bruck, D., Ciolli Mattioli, C., Franke, V., Imami, K., McShane, E., Moreno-Estelles, M., Calviello, L., Filipchyk, A., Peguero-Sanchez, E., et al. (2017). RNA localization is a key determinant of neurite-enriched proteome. Nat Commun 8, 583.

Selbstständigkeitserklärung

Hiermit erkläre ich, Andrei Filipchyk, die vorliegende Dissertation selbstständig verfasst und dabei keine anderen als die angegebenen Quellen und Hilfsmittel benutzt zu haben.

Andrei Filipchyk,

Berlin, 15. Januar 2018